Analytic Philosophy

John-Michael Kuczynski

ANALYTIC PHILOSOPHY

J.-M. Kuczynski, PhD

TABLE OF CONTENTS

PART I
The Analysis of Analysis
Chapter 1
Analytic Philosophy as Logical Analysis
Chapter 2
Properties and Non-Spatiotemporal Existence
Chapter 3
What Are Propositions and What Is Truth?
PART II
Language and Thought
Chapter 4
What is Language?
Chapter 5
Do We Think in Words?
Chapter 6
Russell's Improvements on Frege's Work
Chapter 7
Logic, logicism, and the viability of attempts to mechanize thought

Chapter 8
The Relationship between Thought and Language:
Literal Meaning vs. Cognitive Content
Chapter 9
Putnam's Insight and Burge's Blunder:
Semantic Externalism vs. Content-Externalism
PART III
Knowledge and Inductive Inference
Chapter 10
Knowledge
Chapter 11
Cartesian Skepticism and the Birth of Epistemology
Chapter 12
Skepticism and the Justification of Inductive Inference
Chapter 13
Empiricism and Its Limits
PART IV
Necessity, Causality, and Personal Identity
Chapter 14
Determinism, Randomness, and Unpredictability
Chapter 15
The Compatibilist Conception of Freedom
Chapter 16
Personal and Objectual Identity
Chapter 17

Causality
Chapter 18
Modality and Non-Existence
PART V
Ethics and Religion
Chapter 19
Some Fundamental Principles Relating to Ethics
Chapter 20
Emotivism
Chapter 21
Moral Conventionalism and Moral Nihilism
Chapter 22
The Nature of Subjecthood and the Connection between Self-Interest and Morality
Chapter 23
Kant’s Ethics and His Attempt to Identify Morality with Self-Interest
Chapter 24
Hedonism, Egoism, Utilitarianism, and Deontology
Chapter 25
Religion
Chapter 26
Existentialism
Chapter 27
Law

Appendix 1
The Rudiments of Logic
Appendix 2
Important Terms and Principles

PART I

The Analysis of Analysis

Chapter 1

Analytic Philosophy as Logical Analysis

1.0 Philosophy as the analysis of the categories in terms of which understand the world

We understand the world in terms of certain categories (e.g., *person, statement, fact, impossibility, existence*). Philosophy studies these categories. It delineates their structures. This is its sole function. Thus, *philosophy is the discipline that delineates the structures of the categories in terms of which we think about the world.*

We've already identified some of these categories. By stating a few of our most basic beliefs, we can identify some more (they're the ones denoted by the italicized expressions):

> The world isn't homogeneous. It is articulated into *events* bearing various *causal* and, more generally, *spatiotemporal* relations to one another. Many of these events involve more or less persistent *things* (rocks, trees, etc.). Some of these things have *minds.* Most animate (mind-having) beings have *sense-perceptions;* i.e., they see things, hear things, etc. Most percipient (perception-having) creatures have *beliefs.* Some of these

beliefs are *true;* others are *false.* Some of the true ones are cases of *knowledge.* Some percipient creatures communicate with one another through the use of *language.* Mastery of a language makes it easier for one to communicate one's beliefs to others, and it also enhances one's ability to *reason.* Rational (reason-capable) creatures tend to make *value-judgments.* They judge one another's actions, and sometimes their own, to be *good* or *bad.* Creatures that make such judgments tend to regulate their behavior towards one another by means of systems of *law,* the supposed purpose of which is to ensure that such behavior satisfy the requirements of *justice.*

1.1 Knowledge vs. meta-knowledge

Event, space, time, cause, persistence, thing, mind, perception, belief, knowledge, language, truth, value-judgment, law, justice. Whatever we know, we know it through these categories (and others like them—the list is not complete). But even though we cannot think without them, they are seldom the *objects* of thought. The result is that, although we're adept at using them, we know little about them.

For example, we are excellent at distinguishing linguistic from non-linguistic behavior, which suggests that, at some level, we know what conditions a creature's behavior must meet if it is to embody knowledge of a language. But when asked to make these conditions explicit, we find we can do so only with great difficulty and only with partial success. So even though we are good at knowing who knows a language and who does not, we don't know what it is that we know by virtue of knowing this.

The same thing *mutatis mutandis*[1] is true of each one of these categories. We're excellent at distinguishing between moral and immoral behavior. We know that rape is immoral and that donating money to charity (for selfless reasons) is moral. But when asked to identify the principles embodied in these pedestrian and uncontroversial judgments, we have trouble producing theories that don't distort them.

In general, it's hard to identify the principles that guide our thoughts. Self-understanding isn't the mind's primary function. Nor could it be. The

idea of a mind that thinks about nothing other than itself is an ***incoherent***[2] one. Thus, any case of self-awareness, and therefore of self-understanding, is necessarily derivative of, and for that reason of lesser quality than, some other more fundamental sort of understanding.

Being the discipline whose purpose is to delineate the structures of these categories, philosophy has the very non-trivial job of identifying, in as clear and explicit a manner as possible, the conditions that a given thing must satisfy if it is to fall into a given one of these categories. So far as philosophy succeeds in this endeavor, it makes it clear what it is that we are doing when we are doing anything cognitive, be it making an observation or engaging in an extended piece of abstract reasoning. Philosophy is the analysis of the preconditions of all knowledge. It is the analysis of analysis, the logic of logic, the science of science.

1.2 The relationship of philosophy to other disciplines

The philosopher is interested in the *laws governing the laws.* He doesn't want to know what *in fact* holds. He wants to know what it would even make sense to claim to hold. He wants to know the laws that the laws can't break.

Philosophical knowledge is metaknowledge—knowledge about knowledge. The non-philosopher wants to know specifics. What happened? When did it happen? What did it cause? How did it cause it? The philosopher is interested in these questions *only* to the extent that knowing the answers to them helps him understand the categories (*cause, place, time,* etc.) underlying such knowledge.

The stranded motorist wants his car to work. He doesn't care what will get it to do so. The engineer *is* interested in this. But the engineer is not entirely innocent of the motorist's epistemic parochialism. The engineer has no interest in what the laws of physics are *except* in so far as he must know them to create the right mechanisms. But the physicist wants knowledge of the mechanisms only to the extent it will give him knowledge of the laws embodied therein.[3]

The philosopher is to the physicist what the physicist is to the engineer

and, therefore, what the engineer is to the stranded motorist. The philosopher wants to know what causes what and what mechanisms were involved *only* to the extent that knowing this helps him understand *what it is* for one thing to cause another— *only,* that is to say, in so far as it helps him know what it is that one knows in knowing that one thing made another happen.

1.3 How is analytic philosophy different from non-analytic philosophy?

John Stuart Mill (1806–1873), the great philosopher and economist, said that he was an expert in but one science, that being the science of science. What he meant was obviously similar to what we've been saying. And it was similar, therefore, to what Gottlob Frege (1848–1925), the great philosopher and mathematical logician, meant when he said that logic studies not the laws of nature but the "laws of the laws of nature." Whether Frege was right depends on what exactly one means by "logic." But if, by "logic," he meant "philosophy," then Frege's dictum was spot-on.

Frege is often described as the first analytic philosopher. Michael Dummett (1925–), an exceptionally capable contemporary philosopher of language, said that "analytic philosophy is post-Fregean philosophy."[4] What does Dummett mean?

With some exceptions, pre-Fregean philosophers thought they were studying the most general features of the actual world. They thought that, like the botanist, they were in the business of saying how the world is, the only difference between their work and the botanist's being that theirs is concerned with more general features of reality than the botanist's.

Frege showed that this is wrong. Any interest that botanists have in plants that *might* exist, but don't, is subordinate to their interest in what plants *actually* exist. Like all scientists, they are interested in what there *could* be only to the extent that it helps them figure out what there *is.* But with philosophers, it's the other way around. Any interest they have in the actual is subordinate to their interest in the possible.

Unlike non-analytic philosophers, *analytic* philosophers figure out what there could be by analyzing statements. Statements that make sense are those

that an be true, and statements that don't are those that can't. Therefore, statements that make sense describe possible realities and those that don't don't.[5]

But what exactly did Frege do?

2.0 Frege's key insight: Logical form ≠ grammatical form

Frege's legacy to philosophy can be summed up thus:

> (FL)[6] When people have an obviously correct belief that seems to have an absurd consequence, they should ask themselves whether that absurdity really *is* a consequence of that belief. But they frequently don't. Instead they accept the absurdity and, in order to make this mistake of theirs work, they develop *ad hoc* hypotheses as to the nature of reality that undermine the integrity of their own belief system.

Consider the statement:

(SC) nothing is a square circle.

SC is true. Everybody knows this. But what does SC say? Judging by its grammatical similarity to "Smith is a very capable lawyer," which attributes a certain property (that of being a capable lawyer) to a certain individual (Smith), SC would seem to say that a certain entity has the property of being a square circle. Presumably, the entity in question is some non-entity. If this presumption is correct, SC says that:

(SC*) some non-entity, some featureless un-thing, is a circle.

But SC* is doubly incoherent. If *anything* is a square circle, then SC is false—it being irrelevant how much of a cipher the entity in question is. Second, the very idea of a non-entity is an absurd one. But it's hard to find a layperson or scholar who, when asked what SC means, comes up with anything substantively different from SC*.

Given what a rank absurdity SC* is, we can't accept it, even though it seems to be an obvious consequence of SC. But we can't reject SC, since

it's an obvious truth.

To get out of this jam, we need only reflect a bit on what our words mean. If you say that nobody likes Larry, you're not saying that some un-person likes Larry. You're saying that if you gathered together all the people who liked Larry and put them in an otherwise empty room, that room would remain empty. Which is the same as saying that, if you put all the people in existence in an otherwise empty room, there would be no things that liked Larry in that room.

Thus, what "nobody likes Larry" means is not that some un-person likes Larry, and is instead that the set of people who like Larry is empty or, alternatively, that the set of people doesn't overlap with the set of things that like Larry.

SC is to be understood along similar lines. If you put all the square circles in existence in an otherwise empty room, that room would remain empty. This is another way of saying that, if you put all the square things in an otherwise empty room, there would be no circles in that room. So SC says, not some un-thing is a circle, but that

(i) the set of things that are both circular and square is an empty one

or, alternatively, that

(ii) the set of circles doesn't overlap with the set of squares.

(i) is a way of saying that any given thing lacks the property of being both a circle and a square. Alternatively, it's a way of saying this property doesn't have any instances—that it's *uninstantiated.* (ii) is a way of saying that anything having the property of being a square lacks the property of being a circle. Alternatively, it's a way of saying that these two properties don't have any instances in common—that they are not *coinstantiated.*

(i) and (ii) thus attribute properties to properties. They say, respectively, that the property of being both a square and a circle is uninstantiated and that the properties of being a square and of being a circle are not coinstantiated.

Since (i) and (ii) are just different ways of saying what SC says, the latter doesn't make the absurd statement that some non-entity is a circle, and it instead makes the innocuous statement that the set of squares doesn't overlap

with the set of circles or, alternatively, that the property of being a square circle is uninstantiated.

What this shows is that, in at least some cases, philosophical insight is acquired, not by doing parascience, but by analyzing meanings—not by positing entities, but by clarifying statements.[7]

Analytic philosophy is philosophy that is driven by accep-tance of FL. Let us now expand on FL and make it clear why it's true.

2.1 The wrong way to react to grammatical surface-structure

In respect of its grammatical form,

(JS) "John smokes"

is comparable to

(MS) "Mary smokes."

JS attributes the property of being a smoker to John; MS attributes that property to Mary.

In respect of its grammatical form, each of JS and MS is comparable to:

(LJ) "Larry juggles"

and also to

(JJ) "Jane jogs."

Each of these sentences says of some individual that he or she has a certain property. The obvious inference to make is that *any* sentence grammatically comparable to any one of those of four sentences says of some individual that he or she (or it) has some property.

In light of these points, consider the sentence:

(SS) Someone smokes.

Given what we just said, the obvious thing to say about it is that it attributes the property of smoking to some individual. But which individual

could that be? Which individual does "someone" pick out?

"It picks out an ambiguous person," said one logician.[8] But this answer is no good. Words are ambiguous, not people. "Bank" is ambiguous, since it has two meanings. But I'm not ambiguous, and neither are you. And if *per impossibile* there *did* exist some ambiguous person—some blank, featureless shell of a person who was picked out by "someone"—SS would unambiguously say of *that* person that he or she smoked. But there clearly isn't any *one* person to whom SS attributes the property of smoking. This is easily shown.[9] If Smith smokes, the sentence: "someone smokes but Smith does not" will be *false.* But it won't be ***self-contradictory;*** it won't be like "Smith smokes but Smith does not smoke." (For a statement to be self-contradictory is for it to bear two mutually opposed meanings.) Of course, there isn't anything special about the name "Smith," what we just said could have been said in connection with Jones or Brown or any other expression that refers to some individual. Thus, there is no individual N such that it is self-contradictory to say that someone smokes but N does not. Therefore "someone" doesn't refer to anyone.

"But you've misunderstood my thesis" it will be said. "The word "someone" doesn't unambiguously pick out an ambiguous person. It is *itself* ambiguous. It refers to Fred and Ethel and Mary. It refers to all people indifferently."

That's false. "Someone" isn't ambiguous; it isn't like the word "dumb." SS has one meaning, unlike "John is dumb," which could mean either "John is unintelligent" or "John is mute." Also, if "someone" were *ambiguous* between "John" and "Ethel" and so on, then, depending on the circumstances, it would be synonymous with "John smokes" or "Ethel smokes." And in that case, "someone smokes, but John does not" would sometimes have the same meaning as "John smokes but John does not smoke," in which case it would be self-contradictory. But, as we just saw, "someone smokes, but John does not" is not self-contradictory under any circumstances. If John *does* smoke, it will be *false,* but it won't be self-contradictory.

Also, it isn't clear what it means to say that "someone" picks out everyone "indifferently." But if, as is surely the case, picking out everyone indifferently involves picking out everyone, then it's just wrong to say that

"someone" picks out everyone "indifferently" (or in any other way). For if it did, it would have the same meaning as "everyone," which it doesn't.

2.2 The right way to react to grammatical surface-structure

What all this shows is that "someone" isn't in the same category as "John" and "Ethel." It doesn't function in the same way, even though, given its *grammatical* function, one would expect it to. SS is obviously true. The wrong way to react to that fact is to twist reality to make it conform to our assumption that "someone," being grammatically comparable to "John," must refer to something. If we take that path, we must say that it refers to an "ambiguous" or "non-specific" person, or some such, given that it obviously doesn't refer to anyone specific. But then we're then stuck with the absurd thesis that there is some non-specific entity in the world— that there exists something that isn't identical with any particular thing and therefore isn't identical with anything.

The right way to react to it is to think more deeply about what SS is really saying. Frege did this, and he solved the puzzle. SS's ***logical form*** diverges from its *grammatical* form. In terms of what its grammar suggests that it is saying, SS is indistinguishable from sentences that attribute the property of smoking to specific objects. In terms of what it really is saying, it is evidently very different from such sentences.

What SS is saying, as Frege made clear, is that the characteristic—or as analytic philosophers put it, the ***property***—of being a smoker is instantiated. SS is making a statement, not about some non-specific individual, but about a very specific property, and it's saying of this property that it's ***instantiated.*** (For a property to be instantiated is for there to be an ***instance*** of it. An instance of a property is something that has it. You are an instance of, and therefore instantiate, the property of being human since, being human, you have that property.)

There are properties that nothing has. Nobody has run a three-minute mile. Given any individual, x, it is false to attribute the property of being a person who has run a three-minute mile. Thus, there are no instances of that property. It is uninstantiated. The grammatical form of ("TM" can be thought

of as short for “three minutes”):

(TM) “nobody has run a three-minute mile”

is just like that of

(JTM) “John has run a three-minute mile.”

JTM clearly attributes that property to an individual. That isn’t what TM does. TM says of that property that it *can’t* be attributed to anyone. So TM says that (“UP” is short for “uninstantiated property”):

(UP) the property of being a person who has run a three-minute mile is uninstantiated.

Notice that UP’s grammatical form is the opposite of TM’s. The grammatical subject of TM is “nobody”; the grammatical predicate is “has run a three-minute mile.” By contrast, the grammatical subject of UP is “the property of being a person who has run a three-minute mile,” which corresponds to “has run a three-minute mile,” and the grammatical predicate of UP is uninstantiated, which corresponds to “nobody.”

If one were to take TM at face value, and were thus to assume that its meaning paralleled that of JTM, one would have to say that it attributed the property of being a person who has run a three-minute mile to some un-person, or some such. But if this is what it said, then in order to be true, somebody—namely, this unperson— *would* have had to run a three-minute mile, in which case TM would be false. When we align TM’s real meaning with its grammar, we don’t have to swallow this rank absurdity. For UP says of some very much existent property (that of being a person who has run a three-minute mile) that *it* has a certain very much existent property (that of being uninstantiated). So we get the right result if we do linguistic analysis. We get the wrong result if we do para-science—if, that is, we posit new *entities* in order to account for the datum that TM is true.

The same thing is true of:

(NS) “nothing smokes.”

Pre-Fregeans said that NS says of some non-thing—some blank entity—that it smokes. So, supposing that in 500 years nobody smokes, and that NS is

therefore true, it will be in virtue of the fact that some blank entity is around that is smoking.

But if there is such a thing, and it's smoking, then NS is *false.* As long as one thing smokes, NS is false. It doesn't matter how blank or otherwise deficient that thing is.

Frege straightened this all out. By obvious extensions of what we just said, NS's real meaning is:

(NS_2) *the property of being a smoker* **is instantiated.**

NS_2's grammatical subject (the italicized expression) corresponds to NS's grammatical predicate; and NS_2's grammatical predicate (the bold-faced expression) corresponds to NS's grammatical subject. So when we align NS's logical form (what it actually says) with its grammatical form (what, given its grammar, it appears to say), we no longer have to say, completely absurdly, that NS attributes to some non-specific individual.

Similar remarks hold with respect to SC. That statement says of two properties that they aren't coinstantiated. It doesn't say of some non-entity that it is both a square and a circle.

Philosophical puzzles are solved by making it clear what statements mean. This is the basic tenet of analytic philosophy. Philosophy explains by clarifying statements. Science explains by positing entities. Science posits entities that are not themselves directly encountered but that, if assumed to exist, would account for phenomena that *are* directly encountered.

It used to be thought that philosophical explanation was to be understood in the same way—that philosophical progress was to be made by positing entities that are not themselves directly known but that, if assumed to exist, explain what *is* directly known. This is not the case. Philosophy isn't para-science. Philosophy is conceptual analysis. To make a philosophical discovery is not to discover a new entity; it is to make explicit a previously unrecognized implication of an existing belief. Philosophy is explication; it is the clarification of the statements that we accept but whose depths we haven't yet fully fathomed.[10]

3.0 Analysis vs. ontogenesis

In the works of analytic philosophers, one sometimes comes across the word "ontology," which, etymologically, means "the study of being" ("ontos"= "being," "logos"= "study"). These days the word "ontology" is usually used to refer to a given philosopher's beliefs as to what exists. Some philosophers don't believe in nonspatiotemporal entities. So such entities don't belong to their "ontology." Since I do believe in them, they do belong to my "ontology."

Analytic philosophers are, almost by definition, ontologically very conservative. In other words, they don't want to grant existence to anything whose existence hasn't been demonstrated beyond a shadow of a doubt.

Pre-analytic, pre-Fregean philosophers were ontologically very liberal. Consider the statement:

(TP[11]) no person is over 20-feet tall.

Pre-Fregean philosophers took TP to say that there existed some non-person who was over 20-feet tall, and they engaged in a great deal of spurious "ontologizing" to validate this analysis. Frege showed that this ontologizing, in addition to being futile, is unnecessary. TP says that a certain property (that of being a person who is over 20-feet tall) has another property (that of not being instantiated). No 20-foot tall (un)person need be postulated. No ontologizing need be done. By contrast, non-analytic (pre-Fregean) philosophers often countenanced bizarre and even logically impossible entities to cover up deficits in their analyses.

3.1 When is it appropriate to ontologize?

As a general rule, analytic philosophers do not try to solve philosophical problems by "ontologizing"—that is, by positing an entity, or class of entities, not previously believed to exist. They try to solve them by clarifying statements. In *some* cases, the results of a successfully clarified statement *do* demand that we grant existence to something to which we'd otherwise deny existence. But, whenever this happens, the "entity" being posted is *never* a denizen of the spatiotemporal world; the thing that's posited is never a person

or a table or a mountain or a monster. It's always an abstract object of some kind, and it's posited *only* because, were it not to exist, it would be impossible to account for the truth of obviously true statements.

For example, given the premise that Bob and Sally are both humans who are intelligent, it follows that there exist characteristics—or, to use the word preferred by analytic philosophers, ***properties***[12]—that Bob and Sally have in common, and from this it follows that properties exist.

Given that properties exist, are they identical with spatiotemporal entities (i.e., with things that have locations in space and time)? Obviously, *instances* of properties at least sometimes exist in space-time. Bob and Sally, both instances of many properties, exist in space and time. But although you may encounter many instances of intelligence, you'll never encounter intelligence *per se* and it would make no sense to assign any spatiotemporal location to it. Attempts to rebut this argument are doomed to fail, as we'll see in Chapter 2. Thus, to validate the rather rudimentary inference from "Bob and Sally are both intelligent and kind," it is necessary to grant the existence of non-spatiotemporal entities and thus to do a bit of ontologizing.

So even though analytic philosophers do ontologize, they do it *only* when there is no other way of demonstrating the legitimacy of some indisputably correct form of inference, and they never posit anything nonspatiotemporal in the process.

3.2 Brentano and Meinong: the non-analytic method epitomized

According to Franz Brentano (1838–1917), the essence of the mental is *intentionality.* In this context, the word "intentionality" refers, not to the property of being done deliberately, but to the property of being *representational.* So, in Brentano's view, for something to be a mental entity is for it to be representational, and for something to fail to be a mental entity is for it to be non-representational.

To be sure, there are non-mental things (e.g., deposits of ink or pain) that are representational. Utterances and ink deposits are representational, even though they aren't mental. But this isn't a threat to Brentano's position.

Utterances (etc.) are representational only in a derivative sense. It's because *we* endow it with meaning that an utterance of "snow is white" is meaningful; in a world devoid of sentient beings, it would just be another noise. So Brentano's thesis is that for something to be mental is for it to be non-derivatively representational.

There is an apparent problem with Brentano's thesis. To be representational is presumably to represent something. Hallucinations are mental entities. But what does a hallucination of a pink elephant represent? A pink elephant? No—pink elephants don't exist. More formally, there is nothing x such that x is a pink elephant. *A fortiori* there is nothing x such that x is a pink elephant that is represented by some hallucination. Still, there is clearly *a* sense in which hallucinations of pink elephants and other non-entities are representational. How is this to be explained?

The solution to this problem lies in the fact that *perception is description.* If you're looking at an actual elephant, which we'll call "Larry," the information encoded in your visual ***perception*** isn't the effect that:

(LP[13]) *Larry* is standing over there, next to that tree, looking ill [etc.].

When you look at an elephant, or a person, or a rock, you don't *just* see that object. Seeing Larry involves seeing *a* thing having various properties—having a certain color, shape, size, position (relative to you), etc. So seeing Larry involves seeing that various properties are instantiated—that there is an instance, in a certain place, of a certain morphology, color, etc.

Perceptions of things aren't like sentences about them. In the *sentence:*

(LS[14]) "*Larry* is standing over there, next to that tree, looking ill [etc.]"

Larry is represented by a single, ***semantically simple*** symbol (namely, "Larry").[15] But in no *sense*-perception of Larry is he represented by some simple, homogenous, non-composite cipher. So far as he is perceived, Larry is represented *as* having these or those properties. This means that seeing him involves seeing that these or those properties are instantiated.

It's not as though *in addition* to seeing an instance of a certain morphology, color, etc., you *also,* separately from that, see Larry.[16] No,

your seeing Larry *consists* in your seeing those property-instances—in your seeing that such and such properties are instantiated. Thus, the content of your perception of Larry is given by some ***existence-claim*** similar to the following:

(LC) there exists, over in that place at the present time, instances of such and such properties.

An existence-claim is any claim to the effect that some property is instantiated. Thus, "there are prime numbers" is an existence-claim, and sufficient since it says that the property of being a prime number has at least one instance.

Given any existence claim, anything having the requisite properties is said to *satisfy* it. So the number seven satisfies the just-mentioned existence-claim. The number two *uniquely* satisfies "there is an even prime," since (a) it satisfies it and (b) nothing else does.

When you look at Larry, your eyes are giving you an existence-claim. Given that Larry, and Larry alone, satisfies that existence-claim, he is the *object* of your perception. So given that Larry uniquely instantiates the property of being a thing in such and such a place that has such and such a morphology (etc.), Larry is the object (or, more likely, one of the objects) of your current perception. So, yes, your perception does represent Larry. But it represents him by way of ***encoding*** an existence-claim that he satisfies.

Given this, suppose that, the next day, you have a hallucination that is experientially just like the ***veridical*** perception we've been talking about. (A "veridical" perception is an accurate one. "Veridical" is to perceptions what "true" is to sentences. For some reason, perceptions are described as "veridical," not as "true.") So even though neither Larry nor anything that looks like him is in front of you, your visual experience is telling you otherwise. That hallucination thus gives you a *false* message. The message encoded in it will be similar to LC. That message is to the effect that, in a certain place, there is a thing having such and such morphological, chromatic, kinematic (etc.) properties. On this occasion, the message—the existence-claim—in question is false, the reason being that nothing satisfies it. (Yesterday, the message represented by your visual experience was correct, the reason being that the existence-claim in question *was* satisfied.) The

important point is that, although it was hallucinatory, your visual experience encoded an existence-claim and, in so doing, gave you a message and was therefore representational. It wasn't representational in the sense that there was some thing that it picked out and, for that reason, represented. It was representational in the sense that it gave you a message, albeit a false one, and thus represented the world as being a certain way.

Your visual perception of yesterday, unlike your visual perception of today, encoded a *true* existence-claim. But your visual experience of today no more represents some non-existent entity than your sense-perception yesterday. Your visual experience today has for its content a *false* proposition to the effect that *there is* a thing having thus and such properties. But just as the sentence "there does not exist a ten-foot tall man" does not, in order to be true, require the existence of a ten-foot tall man, so your perception doesn't require the existence of an elephant before you.

The same thing is true of thoughts about the non-existent. When you think about some non-existent number—for example, an even prime greater than two—there isn't some mathematically impossible entity that you're cognitively locking onto. What's going on is that you're thinking some false existence-claim along the lines of: there is some number n such that n is both even and prime [etc.].

Mental entities have **propositions** for their contents. When correct, those propositions *describe* existing things. When false, they don't. But nothing non-existent or quasi-existent can be the object of a thought or perception. When we *describe* a thought or perception as having a "non-existent object," what we mean, so far as what we mean is coherent, is that it has for its content an existence-claim that nothing satisfies.

But Brentano dropped the ball. He realized that hallucinations are, in some significant sense, representational. Wishing to reconcile this with the fact that there are no pink elephants (etc.), he said that a hallucination of a pink elephant has a non-existent pink elephant for its object.[17]

But that's absurd, since it's the same as saying: "there exists some elephant x such that x doesn't exist and such that what you are hallucinating is x."

In a failed to attempt to deal with this, Brentano distinguished different kinds of non-existence, and he used different terms to mark them ("inexistent," "non-existent," "un-existent," etc.) His pupil, Alexius Meinong (1853–1920), added another bogus category to this list—the category of "subsistent" entities. An entity "subsists" if it doesn't quite fail to exist, but doesn't quite succeed in existing either.

This entire approach is misguided. Brentano and Meinong were ontologizing when they should have been analyzing. Properly analyzed, hallucinations no more require the existence of non-existent existents than veridical perceptions. The same is true of thoughts about Bart Simpson, the Fountain of Youth, etc. (This is further discussed in Chapters 6, 8, 9, and 25.) [18]

4.0 Empirical puzzles vs. philosophical puzzles

Some puzzles result from ignorance of spatiotemporal facts. My valuables start disappearing. I'm puzzled. I learn the relevant facts: Larry has been sneaking into my house and stealing my valuables. I'm no longer puzzled.

The puzzles that science deals with typically involve a failure to know all the facts. The problem isn't that anyone is making erroneous inferences. The problem is that not all the facts are in. It isn't yet known that disease X results from an over-production of antibody Y. The reason it isn't yet known is that, given the available data, there isn't yet good reason to believe it. But once the data is in, it will be believed, and a cure will be forthcoming.

Of course, scientific puzzles seldom result *entirely* from a failure to have the relevant data. In most cases, scientific breakthroughs involve somebody's figuring out a new and better way to model already available data.

The pre-history of relativity theory vividly illustrates this principle. If a train rushes past you at a rate of 100 mph, and I rush past you at a rate of 70 mph, the train is traveling at a rate of 30 mph with respect to me. But if a light beam rushes past you at a rate of 186,000 miles/second, and I rush past you at a rate of 180,000 miles/second, the light beam rushes past *me* at a rate of 186,000 miles/second. There is thus is no optical test for determining one's

own state of motion.[19] In other words, no matter how quickly you travel, you will not be able to detect any change in your velocity *relative* to that of a light beam. People and instruments not traveling with you will be able to detect changes in *your* velocity relative to that of a light beam, and you will be able to detect changes in *their* velocities relative to that of a light beam. But nobody can detect any changes in his own velocity relative to that of a light beam.

This deeply puzzling fact was established in 1879. But nobody had any idea how to explain it until, in 1905, Einstein put forth the Special Theory of Relativity. Einstein didn't cite any data that hadn't been available to the physics community for decades. Einstein's great innovation was of a conceptual nature. The facts were in, but he was the first to make sense of them.

Be all of this as it may, Relativity Theory is an ***empirical*** theory. It's based on observational and experimental data, much of which would be impossible to acquire except through carefully executed experiments. This data wasn't in until 1879. So even though nobody came up with Relativity Theory before 1879, that fact can't be chalked up to the fact that nobody drew the right inferences. It is due, at least in part, to the fact that the necessary data simply wasn't available. (What *can* be chalked up to a failure to draw the right inferences is the fact that, during the period from 1879 to 1905, nobody came up with Relativity Theory.)

In general, scientific breakthroughs have two components: (i) a strictly fact-based component (new raw data is acquired), and (ii) a purely conceptual component (already known data is modeled in a new and better way).

4.1 Empirical puzzles vs. philosophical puzzles (continued)

Unlike scientific puzzles, philosophical puzzles are not solved by generating new raw data. Philosophical puzzles are *purely* conceptual in nature and have no strictly factual component.[20] Philosophical puzzles result, not from a failure to know the facts, but from a failure to draw the right inferences. They result, not from ignorance, but from confusion. In figuring out that:

(NS_2) *the property of being a smoker* **is instantiated**

is what is meant by

(NS) "nothing smokes,"

Frege solved a number of outstanding philosophical problems. But Frege didn't make any new empirical discoveries. In fact, his work didn't involve him having access to *any* empirical information that wasn't available to *anyone* who knows what (NS) (or its German equivalent) means.[21] Einstein's work, by contrast, had a heavy empirical component: he was modeling facts that a layperson would know nothing about, and the same thing *mutatis mutandis* is true of any scientific discovery.

4.2. Philosophical analysis ≠ linguistic analysis

Impressed by Frege's philosophical successes, many came to the conclusion that *all* philosophical puzzles are of a purely linguistic nature. The most famous, and also the most vehement, advocate of this view was ***Ludwig Wittgenstein (1889–1951).*** This thesis was the cornerstone of this work. He urged acceptance of it in practically everything he wrote during his long career. He said it before others said it, and he held onto it long after, for reasons to be described forthwith, most of its erstwhile supporters rejected it. "Philosophical confusion begins when language goes on holiday," he said. Elsewhere[22] he said that all philosophical confusion "lies in a failure to understand the workings of our own language."[23]

But this position is incorrect. There are many philosophical puzzles that don't have anything to do with language, and the solutions to these puzzles cannot be modeled on Frege's solutions to the puzzles discussed a moment ago.

Here's an example from ***epistemology.***[24] You couldn't see the book in front of you were it not for the disturbances of your eyes brought about by the light-rays bouncing off of it. In general, nothing can sense-perceive anything that doesn't affect it. All knowledge of what is in space-time is rooted in sense-perception. (If it's in space-time, it isn't known unless it's directly perceived or evidence of it is perceived.) Taking it for granted that nothing that isn't space-time can possibly be known, many contemporary epistemologists, e.g. Jerry Fodor[25], hold that one can't know of *anything* without being affected by it. Thus, such philosophers hold that:

(JK) John's being aware of the fact that 1 + 1 = 2 involves his being on the receiving end of some causal process initiated by that fact.

I personally regard JK as being absurd in the extreme.[26] The fact that 1 + 1 = 2 isn't comparable to the fact that there is a book in front of you. Were mass-energy distributed differently, there *wouldn't* be a book in front of you. But 1+1 would equal 2 no matter how mass-energy were distributed. Thus, "1 + 1 = 2" says nothing about how mass-energy is distributed, and it therefore says nothing about the spatiotemporal world. The fact it describes must therefore be non-spatiotemporal. Nothing outside of space-time can bear any causal relation to anything. Therefore, JK is wrong.[27]

In any case, the dispute between those who accept JK and those who reject it has nothing to do with language. It's agreed what JK means. What isn't agreed is whether the thing that it's agreed to mean is true.

But the controversies surrounding NS *do* concern language; they concern the semantics of the word "nothing." Those who see "nothing" as being a referring term, like "Socrates," see NS as saying that some featureless entity smokes. Those who deny that "nothing" and "Socrates" belong to the same semantic category don't see NS as saying this. Therein lies the controversy. Nothing comparable to this holds in connection with JK.

Analytic philosophers *do* agree that the right way to figure out whether or not JK is correct is by carefully analyzing its meaning.[28] But they also hold (rightly) that it isn't to be solved through *linguistic* analysis.

It must be emphasized that, according to analytic epistemologists, this puzzle is of a logical, not an empirical, nature. It is to be resolved through statement-analysis *alone,* not through statement-analysis combined with empirical research. They're clearly right about this. JK says that John's knowing that 1+1=2 is ***inconsistent*** *with* his not being on the receiving end of a causal process initiated by that fact. For P to be inconsistent with Q is for it to be impossible for both P and Q to be true. What is impossible or the otherwise non-existent cannot be observed. Thus, observation cannot tell you that anything is inconsistent with anything. So there is no way for it to tell you whether JK is correct.[29]

The philosophy of law provides us with another example of a puzzle that is to be solved through statement-analysis but *not* through linguistic analysis. It's agreed that legal systems *can* be morally good. But it's fiercely debated whether they *have* to be. According to some, a legal system can fail to embody any morality at all. Law is about power, not morality. Advocates of this view are known as ***legal positivists*** (no relation to *logical* positivism). According to anti-positivists, anything that doesn't meet certain minimal standards of morality *ipso facto* isn't a legal system. To be sure, legal systems, like all institutions, presuppose the existence of relatively rigid distributions of power. But if an institution qualifies as a *bona fide* legal system, it is at least partly by virtue of its embodying a certain morality.

Positivists and anti-positivists disagree as to whether:

(LM) "nothing can be a legal system without embodying a certain morality"

is a true sentence. But this debate has nothing to do with semantics. Positivists and anti-positivists are in agreement as to what LM means. What they disagree about is whether the meaning that they agree that it has is a correct one. What analytic philosophers of law, such as positivist H.L.A. Hart (1907–1992) and anti-positivist Ronald Dworkin (1931–) , *do* agree about is that LM is to be resolved on the basis of logical analysis.

Echoing what we said a moment ago, analytic philosophers deny that LM makes an ***empirical*** statement. They're right. LM says that x's being a legal system is inconsistent with x's failing to embody a certain morality. We've already seen why the merits of such a claim cannot possibly be determined on the basis of observation.

No philosophical assertions are empirical. Philosophy analyzes the categories in terms of which we think about the world. It does this by saying exactly what it is that is ruled out by a given thing's falling into a given category—by, for example, a given thing's being an instance of knowledge. Since observation can't tell one whether one statement is inconsistent with another, philosophical assertions are non-empirical.

4.2.1 Not all philosophical analysis linguistic analysis (continued)

Thus, not all philosophical puzzles are linguistic puzzles; and Wittgenstein was wrong to say otherwise. But didn't we ourselves say that philosophical analysis is the analysis of statements? Yes we did, and we were right to do so. To see why, Wittgenstein is nonetheless wrong, we must distinguish sentences from ***propositions.*** Propositions are the things meant by sentences. "Snow is white," "schnee is weiss," and "la neige est blanche" all mean the same thing. There is some one proposition that is the meaning of each of them.[30]

A sentence is true or false depending on whether it has a true or false proposition for its meaning. Thus, when a sentence is described as "true," the property being attributed to it isn't the same as the property that is attributed to a proposition that is so described. For a sentence to be "true" is for it to encode a true proposition. But this isn't what it is for a proposition to be true, since propositions don't encode anything. In Chapter 3, we'll say what exactly it is for a proposition to be true. But the obvious answer, though imprecise, is the right one; namely, for a proposition to be true is for it to fit the facts.

The term "statement" is ambiguous; it has three distinct meanings. Sometimes it refers to propositions, sometimes it refers to the sentences used

to affirm them, and sometimes it refers to the *act* of using a sentence to affirm a proposition. Wittgenstein didn't countenance the existence of propositions, and this obviously had a hand in his erroneously believing that philosophy is the analysis of *sentences,* when the truth is that philosophy is the analysis of *propositions.* Even Frege's analyses of sentences such as "someone snores" and "nothing is a square circle" fail to conform to Wittgenstein's conception of what philosophy is supposed to do, since, as we'll see in a moment, they're analyses of propositions that *involve* analyses of sentences and, therefore, are not themselves analyses of sentences.

One must know at least some English to understand the *sentence:*

(1) "John knows that 1 + 1 = 2"

But the proposition meant by that sentence can be grasped without speaking English; and one needn't know English, or any other given language, to be able to analyze that proposition correctly.

But analyzing (1)—the sentence, not the corresponding proposition —*does* involve such knowledge. A sentence is an expression; it consist of nouns, verbs, etc. Analyzing (1) involves knowing the various grammatical nuances involved in its structure. Analyzing the corresponding proposition has nothing to do with anything relating to grammar or any other aspect of language. No such knowledge is needed to analyze the corresponding proposition.

Even though Frege's work inspired many to identify philosophy with sentence-analysis, Frege himself always made it very clear that propositions are not sentences and that, although sentences are human creations, their meanings are not.[31]

The sentence "the moon is less massive than Earth" is a human artifact; it didn't exist until a few centuries ago. But the truth it expresses is in a different category. That truth exists independently of us. After all, the moon was less massive than the Earth before we came along, and it'll be that way after we sign out. For the same reason *mutatis mutandis,* the falsehood expressed by "the Earth is more massive than the moon" exists independently of our thoughts and deeds, even though that sentence itself is a human artifact.

4.2.2 Not all philosophical analysis linguistic analysis (continued): the nature of sentence-meanings (as opposed to sentences)

What are propositions? They're properties. For a proposition to be true is for the world to be a certain way. The proposition that Smith is in Richmond is true if the world is a certain way and it's false if it isn't. (If mass-energy is distributed one way, Smith is in Richmond; if it's distributed some other way, he isn't.) For a thing to be a certain way is for it to have a certain property. To be round is to be one way; to be square is to be some other way. To be round is to have one property; to be square is to have some other property.[32] Thus, the world's being a certain way is identical with its having a certain property. Since, therefore, the world's being a certain way is identical with some proposition's being true, propositions must be identified with properties and a proposition's being true must be identified with its being instantiated. Propositions are properties and truth is instantiatedness.[33]

It's widely thought that propositions are human creations. This is false. The world was a certain way before we were around; it will be a certain way after we're gone; and the way it is while we're around is up to us only to a limited extent.[34] It follows that, independently of our having any beliefs or, indeed, our doing or thinking anything, certain propositions are true; and it follows from this that propositions are not human creations. Sentences, on the other hand, are human creations. They wouldn't be around if it weren't for us. So sentential analysis is a very different thing from propositional analysis.

4.2.3 Not all philosophical analysis linguistic analysis (continued): Frege's accomplishments reassessed

But didn't we say that Frege's great accomplishment lay in his insights concerning *sentences*—in his seeing that a *sentence's* surface structure sometimes pulled part from its deep structure? And didn't we say that, for

this very reason, Frege was the first *analytic* philosopher—that, as Michael Dummett put it, analytic philosophy is post- Fregean philosophy, the reason being that analytic philosophy is *statement*-analysis? Yes, we did say all this. And yet we just spent a lot of time saying how analytic philosophy is "statement-analysis" *only* in the sense of being statement-*meaning*-analysis; that is, proposition-analysis, as opposed to *sentence*-analysis. How are we to reconcile those various statements with one another?

The term "analytic philosophy" can be construed narrowly or broadly. Construed narrowly, analytic philosophy is post-Fregean philosophy in the sense that it directly flows out of Frege's work. Frege was interested in ***reference, quantification,*** the nature of ***logical truth,*** the extent to which it's possible to ***formalize*** intuitively valid inferences, etc. (These terms will be defined soon enough, if they haven't been already.) The term "analytic philosophy" sometimes refers to what is done by those people who write *about those very questions* and who, in so doing, *are taking what Frege had to say about them into account.* In other words, "analytic philosophy" sometimes refers to the ***philosophy of language,*** along with a related branch of philosophy, known as ***philosophical logic.***[35]

But Frege's work had a profound influence on philosophers who were working in areas that have no direct connection to language or logic. Frege showed that, by thinking clearly, systematically, and self-critically, one could make real headway on philosophical problems in which others had yet to make so much as a dent, despite hundreds of years of trying. "The devil is in the details," as they say. Pre-Fregean philosophers tended to disregard the details. Frege did not. He was a stickler for them. And it was partly, though obviously not entirely, for this reason that he was able to solve problems that his predecessors could not. The word "analytic philosopher" sometimes refers to the sort of philosophy done by people who aspire to approach philosophical problems in the same careful and clear-headed way as Frege.

Thus, there are analytic philosophers of law, analytic ethicists, analytic philosophers of religion, etc. What makes them analytic philosophers isn't that they're talking about language or logic or any of the things that Frege talked about. It's that they believe the problems they're concerned with to be solved in the same coolheaded, logical way that Frege solved problems

relating to language and logic.[36]

But even when taken in the narrow sense, "analytic philosophy" (i.e., philosophical logic/the philosophy of language) is only misleadingly described as the analysis of sentences. Philosophers of language are interested in the sentence:

(NS) "nothing is a square circle"

only to the extent that, by understanding it, they will deepen their insight into concepts of a general kind (e.g., ***meaning, analytic truth, modality***). Linguists, on the other hand, are interested in those concepts only to the extent that an understanding of them will help them understand specific sentences, such as NS. Frege did indeed painstakingly analyze specific sentences. But he did so only because he knew that, by so doing, he could identify general logical principles. Thus, for Frege, his insight that logical and grammatical form pull apart was ultimately just a means to an end, the end being the identification of the actual nature of the bearing- relations that propositions have with respect to one another. So, yes, analytic philosophy is statement-analysis; and, yes, it was Frege's brilliant analyses of *sentences* that availed philosophers of the principles needed to analyze statements properly. But analytic philosophy, even the narrow sense of the term, is not itself sentential analysis.

4.2.4 Not all philosophical analysis linguistic analysis (continued): Wittgenstein—an introduction

Wittgenstein's two best-known works are the *Tractatus Logico-philosophicus* (TLP), which he completed in 1921, and the *Philosophical Investigations,* which he completed in 1949. In many ways, these works are antithetical to each other. But in both of them, Wittgenstein insists that that philosophical problems arise when, and only when, sentences are misused and are solved when, and only when, it is made clear how they are being misused.

In the TLP, Wittgenstein contends that all philosophical problems result from a failure to understand the syntactic rules of the languages we use. Sentences that would, if meaningful, express philosophical propositions are

in all cases ungrammatical nonsense and thus fail to say anything. All such sentences violate the ***syntactic rules*** of the languages to which they belong. (For the time being, "syntactic" may be taken to be synonymous with "grammatical." See Chapter 4, Section 3.3 for a definition of "syntax.") Because they violate these rules in subtle, easily overlooked ways, they aren't always seen for the abject nonsense that they are. But we mustn't let the appearances deceive us. All such sentences are syntactically ill-formed and therefore devoid of meaning, and there would be no philosophical problems if people fully understood the syntactic rules of the languages they spoke. So far as philosophy has any legitimate function, it is to identify these rules, thereby heading off the syntactic blunders that lead to philosophical puzzlement.

In the *Philosophical Investigations,* Wittgenstein says, just as he does in the TLP, that philosophical problems arise when language is misused. But in the *Investigations,* he denies that such misuses involve violations of hidden syntactic rules, and instead says that such misuses consist exclusively in one's using words in ways in which they are not *ordinarily* used. Thus, Wittgenstein's position in the *Investigations* is that all philosophical problems can be quickly and definitively solved by looking at how words are *actually* used and, on the basis of the knowledge thereby obtained, ceasing to use words in deviant ways. What we think of as philosophical puzzles concerning knowledge, logic, and morality are puzzles about the words "knowledge," "logic", and "morality." Those puzzles are created by our using those words in non-standard ways, and they're solved by our ceasing to do so.

4.2.5 Is meaning identical with use?

In a moment, we'll evaluate Wittgenstein's (1922) contention that all philosophical statements are ungrammatical nonsense. Right now, let us consider Wittgenstein's (1958) contention that philosophical puzzles are dissolved by looking at how words are used.

This contention is incoherent on many levels. Given only the acoustical and morphological properties of its spoken and written occurrences, the word "knowledge" could mean anything.[37] So, supposing that it's the concept of knowledge that we wish to learn about, we can't possibly know that the word

"knowledge" is the right word to study *unless* we know that it expresses the right concept. But we can't possibly know *that* unless we have some way of grasping that concept that doesn't involve that word. So we can't even act on Wittgenstein's exhortation that we study knowledge by studying how people use the word "knowledge" *unless* we have some way of grasping the concept of knowledge that *doesn't* involve that word. But if we don't need that word to grasp that concept, we don't need it to it study it; and if we don't need it to study it, then Wittgenstein is simply wrong to say that one must study how it is used to understand the concept it expresses. Given any expression E and any concept C, an obvious extension of this argument shows that one can't learn about C by studying E unless one can grasp C, and can therefore examine it, *without* E's help. Thus, Wittgenstein is simply wrong to contend that one learns about concepts by learning about how the corresponding words are used.

"But you've over-stated what Wittgenstein is claiming," it will be said. "Contrary to what you allege, he wasn't saying that concepts are to be learned about *solely* by looking at how the corresponding words are typically used. He was saying only that knowledge of expression-usage would be a useful adjunct to some other, more important way of knowing about concepts." If that's what Wittgenstein is saying, then he's conceding everything said the preceding paragraph.[38] To grant that there is any expression-independent way of grasping concepts is to grant that expression-usage is to be evaluated *in light of* the very conceptual knowledge that, according to Wittgenstein's thesis, one is supposed to acquire *through* the study of expression-usage.

But there's a problem with our argument. It assumes that, given a meaningful expression (e.g. "knowledge"), there is some object that is its meaning. Wittgenstein rejects this assumption. He holds that there is no *entity* that is the meaning of "knowledge." Wittgenstein's position is, I quote, that "meaning is use."[39] In other words, for an expression to have a given meaning is for it to be used in a certain way. An expression's having a given meaning does *not,* in Wittgenstein's view, involve there being some entity that is its meaning.

Before we evaluate this contention, we must make a few facts about it

clear. Wittgenstein isn't making the uncontroversial point that how expressions are used is a function, in part, of what they mean. Nor is he making the equally innocuous point that how an expression is used may have *effects* on what it means. ("Probable" used to mean "capable of being definitively established." Enough people used it to mean "likely, but not certain"; and, for that reason, that's what it now means.) He is saying that *what it is* for an expression to have a given meaning is for it to be used in a certain way. Here is his argument:

> (WA[40]) According to some philosophers, there is some entity that is the meaning of "snow is white" and some other entity that is the meaning of "grass is green." These entities are known as *propositions*. Propositions (if existent) are non-spatiotemporal entities.[41]
>
> The very idea of such an entity is of doubtful coherence. And even if such entities do exist, they're nothing to us. We can't see them or touch them or otherwise have anything to do with them. So, even if they do exist, they have no role in human affairs. At the same time, "snow is white" and "grass is green" differ in meaning. So we must find a way of saying what it is for two expressions to 'differ in meaning' that doesn't involve our positing meanings.
>
> This can be done. "Snow is white" isn't used in the same way as "grass is green." There are situations that prompt utterances of the one that don't prompt utterances of the other. By the same token, if they *were* used in the same way—i.e., if there were no situation that prompted utterances of the one that didn't prompt utterances of the other—then they wouldn't differ in meaning. If the sensory stimulations that induced people to say "grass is green" coincided with those that prompted people to say "snow is white," there would be no significant sense in which they "differed in meaning."[42] Thus, two sentences S_1 and S_2 coincide in meaning exactly if they are used in the same way. So instead of identifying the meaning of "grass is green" with some non-spatiotemporal entity—with some entity whose existence is in doubt and whose role in human affairs, supposing it to exist, is also in doubt—we can just say that two expressions "have the

> same meaning" if they're used in the same way. An obvious corollary is that for an expression to have this as opposed to that "meaning" is for it to be used in this as opposed to that way.

Were WA cogent, Wittgenstein would be doing to meanings what Frege did to square circles. Frege got rid of square circles by showing that sentences appearing to require their existence (e.g. "square circles simultaneously have, and lack, uniform curvatures") are equivalent to sentences that don't (e.g. "the statement *x is a circle* entails *x has a uniform curvature* and the statement *x is a square* entails *x does not have a uniform curvature*"). Wittgenstein wishes to get rid of meanings by showing that statements appearing to require the existence of expression-meanings are equivalent to statements (about expression-usage) that don't.

This wish of Wittgenstein's embodies some very wrong views as to what linguistic expressions are. A burst of noise that doesn't have a meaning is just a burst of noise. A burst of noise is an *expression* only if it has a meaning. But if it already has a meaning, then how it's being used isn't what gives it its meaning.

Our knowledge of what words mean is what *guides* our linguistic behavior. Expressions have meanings. We know this. And that's why we use expressions in the way we do. I know what "hug" means; I also know what "discuss" means. That's why, when talking to my students, I say "I want to discuss the exam," and not "I want to hug the exam." If meaning were use, meaning couldn't *guide* use. But it obviously does.

To the extent that meaning *doesn't* guide use, people are *mis*speaking. Wanting to tell you that you're an absolute genius, I say "you're an absolute moron," since, despite my generally good command of the English language, I wrongly think that "moron" means *genius.* To the extent that my uttering those words embodied a *failure* to know the actual meaning of what I was saying—to the extent that meaning *failed* to guide use, in other words—I *mis*spoke. And to the extent that my uttering those words embodied a knowledge of what they meant—to the extent that meaning *did* guide use, in other words—I spoke properly. This shows that, to the extent that utterances aren't simply defective, meaning guides use and, consequently, that use is *not* constitutive of meaning.

Bearing these points in mind, let us revisit Wittgenstein's (1958) contention that it's a philosopher's job, not to analyze concepts or meanings or other such alleged phantasms, but merely to take note of *when* words like "justice," "knowledge," and "truth" are used. What would it be to do this? The situations in which the word "justice" is used don't necessarily have any observable characteristics in common with one another. It's not as though people utter the word "justice" when, and only when, they're just been bitten by a cat. This isn't to say that uses of the word "justice" cannot be correlated in *any* way with facts about the situations in which those uses occur. Such correlations clearly can be made. People are likely to use that word when they're in classes concerning **ethics** or the philosophy of *law;* they're likely to use it when they **believe** that they've been **wronged**; they're likely to be used by politicians who want people to believe them to be **worthy** of holding office.

The boldfaced terms express extremely abstract concepts; and utterances of those words are no more capable than utterances of the word "justice"' of being correlated with observable facts about the contexts in which they occur. So, while a person's decision to use the word 'justice' (or "knowledge" or "law," etc.) may obviously have a situational basis, there is no way to identify that basis *except* in terms of the very concepts that, according to Wittgenstein, are to be studied by figuring *when* those terms are used. In other words, any correct generalization as to when words like "justice," "logic," "knowledge" and "morality" are used will *itself* employ the very concepts that, according to Wittgenstein, knowledge of such generalizations is supposed to yield.

These reflections bring us face to face with an incoherence inherent in all of the different variants of the contention that philosophy is sentence-analysis. To speak isn't just to make noises; it's to make noises for the reason that one believes those noises to have certain meanings. I'm with my friend Larry. All of a sudden he starts convulsing and foaming at the mouth. I call 9-1-1 and say (i) "my friend is foaming at the mouth and convulsing; please send help." Why did I choose those words? Why didn't I say (ii) "I like pizza" or (iii) '"giraffes are friendly creatures"? Because I know that (i) has the right meaning and that (ii) and (iii) don't. In general, speaking consists in making noises for the reason that one believes that, given existing semantic

rules, those noises have certain meanings. If somebody makes a noise that he does *not* believe to be assigned a meaning by any semantic rule, he isn't speaking. If, not believing that the noises I'm about to make are assigned any meaning by any semantic rules, I say "blurga dunga blurbo," I am not speaking. I'm just making noise. Maybe the semantic rules of some language *do* assign them a meaning. That doesn't matter. My making that noise wasn't guided by my knowledge of such rules. So I wasn't saying anything.[43] And as we saw earlier, I am *mis*speaking if I misidentify the meaning assigned by existing semantic rules to the noises I am making. So there is no *bona fide* speech where there isn't awareness of semantic rules.

Thus, the analysis of sentences isn't the analysis of noises. (In this context, take references to "noises" to be short for references to anything that can constitute the occurrence of an expression—e.g., hand-movements, patterns of light on a monitor, etc.) In and of themselves, noises aren't speech, as we just saw. It is only when a noise embodies an attempt to follow a semantic rule that it constitutes speech. So the analysis of speech—that is, of spoken sentences—concerns noises *only* to the extent that they embody attempts to follow such rules; and for the same reason *mutatis mutandis,* the analysis of sentences *per se,* as opposed to their spoken (or written) occurrences, necessarily involves, if it doesn't coincide with, the analysis of those rules.

Those rules are not themselves sentences. The semantic rule that assigns meaning to "snow is white" is not itself a sentence.[44] Given these points, what might it mean to say that philosophy is sentential analysis? It *could* mean that philosophy is the empirical study of different languages, i.e., that philosophy is linguistics. But that clearly isn't what philosophy is. So it must mean that philosophy is the study of semantic rules *qua* semantic rules—that, in other words, it is the study of the *concept* of a semantic rule.

In that case, philosophy is also the study of those concepts that must be understood to understand the concept of a semantic rule. And there are many such concepts. Among them are narrowly semantic concepts such as ***compositionality, reference, quantification, force,*** and ***negation.***

But many of these concepts aren't *only* of relevance to semantics. As we've seen, no noise constitutes a sentence-utterance unless it embodies an

intention of a certain kind; and there is no linguistic behavior of any kind where people aren't intentionally following what they ***believe*** to be existing semantic ***rules.*** Some believe that semantic rules are ***conventions.*** Others believe that they are ***functions*** (in the mathematical sense)— assignments of meanings or ***truth-conditions*** to noises or to properties that are instantiated by noises (or inkmarks, etc.). There is no way to figure out what semantic rules are without examining these concepts.

And even if these questions are side-stepped, and philosophers focus *only* on narrowly semantic concepts (e.g., reference, compositionality, etc.), philosophy ends up being the analysis, not of sentences, but of *concepts.* Thus, the thesis that philosophy is the study of sentences either collapses into the obviously false thesis that philosophy is the empirical study of language *or* into the thesis that philosophy is conceptual analysis (i.e., the analysis of the concepts in terms of which the world is understood). Thus, Wittgenstein's contention that philosophy is the study of sentences collapses into the very view to which it is meant to be an alternative (viz. that philosophy is the analysis of concepts, as opposed to expressions), and is therefore false.

Let us now turn to Wittgenstein's (1922) Tractarian contention that philosophical statements are ungrammatical nonsense.

4.2.6 The Tractarian contention that philosophical statements syntactically ill-formed nonsense

The sentence

(a) "one can be aware of the fact that 1 + 1 = 2 without being causally affected by that fact"

appears meaningful. So does the sentence

(b) "there could in principle be a legal system that failed to embody any morality at all."

But, in the TLP, Wittgenstein says that that this is an illusion. Here is the viewpoint underlying this bold contention:

(TA[45]) All philosophical statements are ungrammatical nonsense, and all

philosophical problems would vanish if we spoke grammatically. "All philosophical problems belong to the same class as the question whether the good is more or less identical than the beautiful."[46] Thus, all philosophical statements belong to the same class as:

(BG) the good is more or less identical than the beautiful.

BG is obviously meaningless. The reason for this is that it's syntactically ill-formed. If we produced only syntactically well-formed sentences, we wouldn't produce nonsense like BG. Since all philosophical statements are in the same class as BG, we'd never produce any sentences whose merits it was the job of philosophy to determine.

Were the logical forms of (a) and (b) brought into alignment with their logical forms—in other words, were their actual meanings reflected in their grammar—they'd be ungrammatical since they have *no* meanings.

Although TA is a paraphrase, not a quotation, the part in quotes *is* an exact quotation from the TLP.[47]

What Wittgenstein is saying isn't confined to the defensible claim that philosophical problems are to be solved by clarifying statements. Wittgenstein is making two additional claims. First, no sentence that appears to make a philosophical statement means *anything.* Second, it's *only* because they're syntactically ill-formed that such sentences are meaningless—there is no other reason.

Let us evaluate these claims. Though ungrammatical, the sentence:

(MH) "me and Herby play tennis every day, and me always win because Herby not in good shape"

is perfectly meaningful. Ungrammatical statements are often meaningful. Thus, BG's failure to mean anything cannot be entirely blamed on its being ungrammatical.

What's the real problem with BG? The expression "more identical than" is obviously supposed to function in the same way as relational expressions, like "more important than" or "identical with." But there is no relation that it picks out. Though it *consists* of English expressions, the expression "more

identical than" is *itself* no more a part of the English lexicon than "blurga derba gurb." For that reason, the meaninglessness of BG is to be accounted for in the same way as the meaninglessness of:

(BG#) "the beautiful is blurga derba gurb the good."

The problem with BG# is that "blurga derba gurb" doesn't mean anything. It may be that its meaninglessness is *reflected* in its syntax. It may be that *because* "blurga derba gurb" doesn't mean anything, BG#'s syntax is off. But, if so, its syntactic shortcomings are to be explained in terms of its lack of meaning, not *vice versa.*

The same holds of BG. The reason that BG doesn't mean anything is that "more identical than" doesn't mean anything. BG's syntactic shortcomings are to be explained in terms of its lack of meaning, and not *vice versa.*

This is easily verified. Given any relation R, if "more identical than" denoted R, BG would be meaningful. If, for example, it denoted the relation that is in fact picked out by the expression 'a better dancer than," BG would say that the good is a better dancer than the beautiful; and it would thus have the same meaning as

(BG*) "the property of goodness is a better dancer than the property of being beautiful."

BG* isn't meaningless; it's false. The property of goodness can't dance; neither can any other property, including the property of being beautiful. Therefore, the former property isn't a *better* dancer than the latter. Nothing false is meaningless, since to be false is to bear a false meaning. So BG* is meaningful.

Incidentally, according to Gilbert Ryle (1900–1976), a career-long Wittgenstein-hardliner, it cannot meaningfully be said of properties that they can, or cannot, dance.[48] It can be said of a human being that he can, or cannot, dance. But nothing that can meaningfully be said of non-properties, such as human beings, can be meaningfully said of properties themselves.

Ryle's position is false. There are many differences between people and properties. One of them is that people can dance, whereas properties cannot. Ryle's position is also self-refuting. In saying that it cannot be said that

properties cannot dance, Ryle is saying exactly what it is that, according to his theory, cannot be said.

Ryle is confusing absurdity with meaninglessness. The sentence "triangles have four sides" makes an absurd, but meaningful, statement. Given that triangles have three sides, not four, it's false. Given that it's false, it's meaningful. "Properties can dance" is absurd and, therefore, false and, therefore meaningful.

There are two kinds of "nonsense." A sentence can be nonsense by failing to have *any* meaning. (BG# is nonsense in this sense.) And a sentence can be nonsense by virtue of having of an absurd meaning. ("Properties can dance" is nonsense in *this* sense.) Ryle doesn't distinguish between these two kinds of nonsense. Neither does Wittgenstein. And Wittgenstein's position, like Ryle's, is self-refuting—and for much the same reason. If Wittgenstein is right to say that all philosophical assertions are ungrammatical nonsense, that very assertion is ungrammatical nonsense and therefore isn't true.[50]

Given how implausible and illogical it is, why on Earth did Wittgenstein hold that philosophical statements are always ungrammatical nonsense? The answer, I believe, is that Wittgenstein thought this view to be the distillation of Frege's groundbreaking philosophical successes. Frege showed that reparsing sentences sometimes solves philosophical problems. By reparsing

(NS) "nothing is a square circle,"

we show that it doesn't attribute the property of being a square circle to some non-entity, thereby dissolving an age-old riddle. Wittgenstein seems to have inferred from these successes of Frege's that *all* philosophical problems are to be solved by reparsing sentences.

This inference is fallacious. Given only that some philosophical problems are to be solved by reparsing sentences, it doesn't follow that they are all to be solved in that way. And we've seen that, indeed, many of them are not to be solved in that way.

Also, there were no cases where Frege's reparsing of a problematic sentence showed it to be ill-formed. For example, in reparsing NS, Frege showed that its logical syntax differed from its apparent syntax. He didn't show that there was anything *wrong* with its logical syntax.

To be sure, philosophers often produce deeply absurd statements. For example, according to pre-Fregean philosophers, NS entails that:

(NS_1) there exists some non-entity that that is *a* square circle.

If NS_1 is right,

(NS_2) there exists some entity that does not exist that is a square a circle.

NS_2 is self-contradictory and therefore absurd. But it's obviously meaningful.[51] If it weren't, it wouldn't contradict itself. (For a sentence to contradict itself is for it to bear two opposed meanings. "Smith is a lawyer and Smith is not a lawyer" because the one ***conjunct***[52] contradicts the other.) But whereas this sentence is explicitly self-contradictory, the self-contradictory sentences that people actually utter are usually only implicitly so.)

In any case, even if philosophical statements are meaningless, they're not meaningless *because* they're ungrammatical. Second, they're not meaningless. The statement that they're meaningless is itself a philosophical statement. That statement is therefore false if it's true. Therefore it is false. (Any statement that entails its own negation is false.[53])

4.3 The Tractarian roots of Logical Positivism

Although one of the TLP's contentions is that all philosophical statements are ungrammatical nonsense, this is not its main contention. The main contention of the TLP is that:

(CT[54]) a sentence is meaningful if, and only if, it is either a ***tautology*** or an ***observation report.***

A "tautology" is a definitional truth (e.g., "fathers are male," "there are three feet in a yard").

An "observation report" is a statement that reports what one's senses have told one (e.g., "I am now seeing a dog," "there is chocolate syrup (or, in any case, a brown discoloration of some kind or other) on Smith's ice-cream," "I

can see your house from here").

CT can be broken down into two claims:

(1) All meaningful non-empirical statements are *tautologies*

and

(2) All meaningful non-tautologous statements are *observation-reports.*

(1) entails that non-empirical disciplines (e.g., philosophy, mathematics) consist of statements that say nothing about anything. (2) entails that anything non-tautologous that cannot be known *strictly* on the basis of what one's senses tell is meaningless.

Wittgenstein's claim that all philosophical statements are ungrammatical nonsense is a ***corollary***[55] of CT. If they're meaningful, philosophical statements, unlike tautologies, are non-trivial. Consider the statement that:

(KC) "knowing a truth doesn't necessarily involve one's being affected by the state affairs described by that truth."

KC is a philosophical statement; and so its negation. Neither statement is a tautology, and neither statement is empirical. If CT is correct, it immediately follows that both KC and its negation are ungrammatical nonsense. Since philosophical assertions are *never* tautologous, CT entails, as Wittgenstein knew, that philosophical assertions are categorically meaningless.

The position that (1) and (2) are both correct is known as *logical positivism* (LP). During the decade or so following the publication of the TLP, and largely because of it, LP was very popular.[56] But (1) and (2) are false. Let us now say why.

4.4. Verificationism and falsificationism

(2) is identical with a doctrine known as ***verificationism.*** According to verificationism a non-tautologous statement is meaningful ***iff*** it's capable of being verified strictly on the basis of sensory observation.[57]

Verificationism is false. The statement "all metal expands when heated"

is meaningful. But it cannot be conclusively verified (i.e., shown to be true) strictly on the basis of observation. No matter how many metal objects you find to expand when heated, it's a possibility that some metal object that you haven't yet considered will fail to do so.[58]

Even though it cannot be verified "all metal expands when heated" can be *falsified* (i.e., shown to be false) strictly on the basis of observation. Advocates of LP saw this and, having for this reason rejected verificationism, accepted a doctrine known as ***falsificationism.*** According to falsificationism, a non-tautologous statement is meaningful if and only if it's capable of being falsified strictly on the basis of observation.[59]

Falsificationism is false. Though obviously meaningful, the statement "there exists a gold ball that weighs exactly 27.13654 lbs" cannot be conclusively falsified, since no matter how many gold balls you consider, it's possible that some gold ball that you haven't considered has that weight.

Falsificationism is really a version of verificationism. According to verificationism, S is meaningful if verifiable. According to falsificationism, S is meaningful if S's negation is verifiable. So given that verificationism fails, it's no surprise that falsificationism does as well.

In light of the failure of falsificationism, advocates of LP decided to soften their views about meaningfulness one more time. This time, they said that a non-tautologous statement is meaningful if possible observations can ***confirm*** it. (P confirms Q if, other things being equal, Q is more likely to be true if P is true than if not-P is true. Other things being equal, Smith is more likely to be wealthy if he wears fancy clothes than if he doesn't wear fancy clothes. Thus, "Smith wears fancy clothes" confirms "Smith is wealthy.") We'll refer to this view as "confirmationalism."

Confirmationalism is equivalent with the position that all meaningful non-tautologous statements are ***empirical.*** An empirical statement is one that, if true, expresses a truth that must be known through observation and that, if false, is the negation of a true empirical statement. "There are trees in Santa Barbara, CA" is a true empirical statement, and "there are no trees in Santa Barbara, CA" is a false one. The negation of an empirical statement is an empirical statement. This is becasue, if it's an empirical question whether or

not S is true, then it's an empirical question whether or not S is false and, therefore, whether or not not-S is true. (There are, as we will see, very few truths, if any, that can be known *strictly* through observation. A statement is empirical if the truth or falsity of it is to be decided on grounds that are at least *partly* observational.)

Henceforth when we refer to "logical positivism" (LP), we will refer to the position that a **statement S is meaningful if and only if (i) S is a tautology (e.g., "sisters are female siblings") or (ii) S is an empirical statement (e.g., "there are trees in Santa Barbara").**[60]

4.5 Logical Positivism Evaluated

Consider the statement:

(1) "triangles are three sided figures."

(1) is plausibly seen as just being true by convention. The same is true of:

(2) "pentagons have five sides."

For argument's sake, we'll grant that, indeed, (1) and (2) are true by convention—that they're definitional truths.[61] Given (1) and (2), it follows that:

(3) If x is the number of sides of an arbitrary triangle and y is the number of sides of an arbitrary pentagon, then w is an even prime iff w is one less than x and three less than y.

But (3) clearly isn't a conventional truth. Though it follows from conventions, (3) is not itself a convention. (3) is thus a non-tautologous, non-empirical truth. It follows that LP is false.

It's possible to have conventions that are ***inconsistent*** with one another. If I stipulate that "x" unambiguously refers to the number two, and I also stipulate that "x" refers to the number of sides of a triangle, my definitions are inconsistent with one another. The internal consistency of conventions is not itself a matter of convention. It isn't an empirical fact that the conventions just described are inconsistent with each other. To say that P is inconsistent with Q is to say that P *must* be false if Q is true. Observation cannot tell you

what *must* be the case; it can tell you, at most, what *is* the case. So it cannot tell you that two linguistic conventions are inconsistent with each other.

Since observation cannot tell you whether or not two statements are inconsistent with each other, it cannot tell you whether or not one statement is a necessary consequence of some other. This is because for Q to be a necessary consequence of P is to P to be inconsistent with the ***negation*** of Q. (The negation of "snow is white" is "snow is not white." In general, the negation of a Q is not-Q.)

When evaluating LP, one must be careful to distinguish sentences from their meanings. Two different sentences can have the same meaning (e.g., “snow is white” and “schnee ist weiss”). The meaning of a sentence is a ***proposition.*** Propositions are not sentences. The thing meant by “snow is white” is not itself a sentence. Some sentences express propositions that are logically true. A proposition is logically true if the laws of logic prohibit its negation from being true. The proposition meant by

(4) “If a given thing is round, then that thing is not a square”

is logically true, since the laws of logic don’t allow round things to be square. And some authors, for this reason, would describe (4) itself, the sentence, as logically true. But what they mean is that, *given* what it is that it means, (4) must be true. So what they are *in fact* describing as logically true is the proposition that (4) couldn’t be false, given what it is that it means.

Thus, it is always propositions, and never sentences, that are logically true. But no *proposition* is true by convention. It’s up to us what our symbols mean. But it’s not up to us whether those meanings are correct. It’s up to us what it is that “the moon is not made of cheese” means. But it isn’t up to us whether that meaning is correct. LP identifies logical truth with conventional truth: truths of logic are sentences that are true by convention. But that’s false. Logical truths are never sentences; they’re always propositions, and propositions are never conventionally true.

Incidentally, (ii) collapses into (i). Linguistic conventions are known empirically. It can be known only through observation that “triangles have three sides” is true. That sentence could mean anything.[62] It could mean that penguins are smarter than humans; and it’s only because you’ve had the

requisite sense-perceptions that you know it not to mean this.

It should be pointed out that logical positivists were unanimous in denying the existence of propositions and of meanings generally. Logical positivists didn't accept the view that for a lecture to concern triangles is for the meaning of that lecture to concern triangles. This is why Rudolph Carnap[63] (1890–1970), an LPhardliner for many years, said that for a lecture to concern triangles is for the word "triangle" to occur in the lecture.

Carnap's view is false. Many a lecture that doesn't contain the word "triangle" concerns triangles. (Think of all the mathematics lectures given in Japanese, Swedish, and Arabic. How often does the word "triangle" occur in them?) And a lecture that contains the word "triangle" isn't necessarily about triangles. Somebody giving a lecture on linguistics may use the word 'triangle' to illustrate some point about phonetics; but in so doing, that person isn't talking about triangles. They're talking, not about triangles, but about the *word* "triangle."

The distinction between the word "triangle" and the corresponding meaning is one that Carnap couldn't countenance without ceasing to be a logical positivist. The very essence of logical positivism is the denial of meaning. For argument's sake, suppose there to exist objects that are the meanings of sentences. (Following convention, we'll refer to these things as "propositions.") Given the existence of propositions, whether a given sentence is meaningful is *not* to be explained in terms of its being either tautologous or confirmable. Rather, a sentence's being tautologous or confirmable is to be explained in terms of its bearing a proposition of a certain kind. A tautologous sentence would be one that had a logically correct proposition for its meaning, and a confirmable sentence would be one that had a confirmable proposition for its meaning. But if there are logically true propositions, then some truths are *ipso facto* not empirical.

We can use words in any way that we like. We can use the words "if Smith has three boats, then Smith has more than one boat" to mean that 1+1=3. But whatever meaning we end up assigning to those words, it's not up to us whether that meaning is correct. And if, as is actually the case, that meaning is of a logical nature, there is *ipso facto* non-empirical truth, an immediate consequence being that LP is wrong.

Carnap's attempt to do away with meanings consisted in his saying (though he did not himself put it this starkly) that the meaning of the word "snow" was that very word, i.e., that words were their own meanings.[64] Given how brazenly wrong this view is, Carnap's attempt to do away with meanings never had many takers.

A much better received attempt to do away with meanings is to be found in a doctrine known as ***conceptual role semantics*** (CRS). According to CRS, two sentences have the same meaning if, and only if, they are *used* in the same way. So "hace mucho calor" is the Spanish translation of "it's hot" *not because* those sentences share a meaning—meanings don't exist, according to CRS—but because the one sentence is used in the same way as the other.[65]

CRS seems to coincide with the Wittgenstein-Grice thesis that "meaning is use." It's thus a mystery why CRS is so popular, given that Grice's coincident position was universally rejected long ago.

In any case, CRS is indefensible. So far as she isn't misspeaking or randomly barking out noises, anyone who says "it's hot out" or "hace much calor," or any other sentence, does so because she knows that existing semantic rules assign a certain proposition to those words and she wishes to put that proposition into words.

One immediate consequence of this is that Carnap's position is false. Another immediate consequence is that what it is for those two sentences to have the same meaning is *not* for them to be "used in the same way."

Also, given any natural language, there are infinitely many sentences belonging to it that have never been used. The thesis that sentences have the same meaning iff they're used in the same way has the absurd consequence that any two sentences that haven't been used before have the same meaning. (If two sentences aren't used at all, they aren't used differently, and are therefore used in the same way, if only in a vacuous sense.)

4.6 Logical Positivism Evaluated (continued)

Let's resume our discussion of Logical Positivism (LP). We've seen that, contrary to what LP says, there are non-empirical truths that have nothing to

do with anyone's linguistic practices. We'll now see that, contrary to what LP says, there are facts about the spatiotemporal world that cannot possibly be known strictly on the basis of sense-perception.

Let NT be the body of assertions jointly constituting Newton's physics.[66] There is no denying that NT is meaningful. But *by itself* NT doesn't make any predictions or otherwise have observable consequences. Physical laws are expressed by ***conditional*** assertions—that is, by statements of the form ‹if P, then Q.›[67] NT doesn't say anything about how this or that physical object will behave. It says how a given object will behave *if* certain conditions are met. NT says, for example, how an object will behave *if* it has a certain mass and is within a certain distance of another body having a certain mass.[68] But NT itself obviously doesn't say that this or that object has this or that mass or is within this or that distance of this or that other specific body. Thus, *taken by itself,* NT isn't confirmable. What *is* confirmable isn't NT, but NT *plus* statements describing specific matters of fact. What is confirmable isn't NT, but some statement of the form t ‹given such and such, NT makes it likely that thus and such.›

But NT is obviously meaningful. This is a problem for LP. To deal with it, advocates of that doctrine proposed that (ii) be replaced with the position that (ii*) a statement S_1 is meaningful if there is some statement S_2, such that *given* S_2, S_1 is confirmable (i.e., capable of being supported by observation).

By this standard, "the nothing nothings" qualifies as meaningful and so does every other nonsense sentence one can think of. Given the statement "if grass is green, then the nothing nothings," anything that confirms "grass is green" confirms "the nothing nothings," and "the nothing nothings" thus qualifies as meaningful.

LP replaced (ii*) with other, similar proposals. But they all ended failing for reasons similar to the one just discussed.[69]

4.7 LP self-defeating

For the reasons just given, it soon became clear that LP was unsalvageable, and soon everybody jumped ship. In fact, it was some of LP's most staunch

proponents who first made it clear what LP's shortcomings were. We'll see this in this chapter when we discuss the brilliant criticisms of LP put forth by Carl Hempel (1905–1997), who was one of LP's first and most ardent advocates.

But erstwhile advocates of LP tended not to see the incoherence that lies at the center of that doctrine. When saying why they rejected LP, they usually cited narrow, technical problems of the sort just discussed. What they didn't do, but what we're about to do, is to say why LP is at its very core a broken and illogical doctrine.[70]

Anything that is true or false is meaningful. Truth implies meaningfulness and so does falsity. Thus, LP is meaningful if it's correct. LP says that any meaningful statement is either a tautology or is empirical. So if LP is correct, it is itself either a tautology or it is empirical.

It isn't a matter of convention that the expression "meaningful sentence" is interchangeable with the expression "sentence that is either a tautology or is empirical." Therefore, LP isn't a tautology.[71]

Since it isn't a tautology, LP is an empirical truth if it isn't false.

But LP isn't an empirical truth. Any attempt to provide an observational basis for any statement presupposes the meaningfulness of that statement and thus presupposes an answer to the question "what conditions must a statement fulfill to be meaningful?" For this reason, the question "what conditions must a statement satisfy if it is to be meaningful?" isn't empirical in nature. It follows that one cannot coherently attempt to find *empirical* grounds for accepting LP, since any attempt to do so *itself* presupposes the meaningfulness of LP. Thus, LP isn't an empirical theory.

Thus, LP is neither an empirical truth nor a tautology. It is thus a counterexample to itself and is therefore false.

Interestingly, in the TLP, Wittgenstein seems to be aware that one cannot coherently say that there can be non-tautologous meaningful statements. Not a single one of the assertions in the TLP is empirical; and not a single one of them is a tautology. This means that, if the TLP's main thesis is correct, the TLP is nonsense. Wittgenstein acknowledges this. For he ends his book by saying that everything that he says in it is meaningless and that those points

ought to be seen, not as truths, but as ladders that one can use to get to the truth but that, once one actually gets there, one must throw away, since they are not themselves truths. Here are the very last words of the TLP:

> My propositions serve as elucidations in the following way: anyone who understands me eventually recognizes them as nonsensical, when he has used them—as steps—to climb up beyond them. (He must, so to speak, throw away the ladder after he has climbed up it.) He must transcend these propositions, and then he will see the world aright. What we cannot speak about we must pass over in silence.[72]

So far as they aren't trivial, these magisterial words are false. To understand something *is* to see its meaning. Therefore, anything that is understood has a meaning. So Wittgenstein's words, if understood, have a meaning, and Wittgenstein is therefore contradicting himself in saying that those who understand his words will see that they're meaningless. Wittgenstein's awe-inspiring injunction that we pass over in silence what we can't speak about involves a similar solecism. To remain silent about something *is* to pass over it in silence. So Wittgenstein is asking, emptily, that we not say anything about what we can't possibly say anything about.

4.7.1 Empiricism self-refuting

These points are easily extended to show that ***empiricism*** is false if it's true and, therefore, that it's false.

Empiricism isn't the claim that

(1) whatever we know *now,* we learned it through sense-perception—but it's possible that at some time in the future we'll acquire knowledge in some other way.

Empiricism is the doctrine that everything that can be known *must* be known through observation, i.e., that

(2) it's inherent in the nature of knowledge that all knowledge be strictly observation-based.

But it cannot be known through observation that (2) is correct. According to (2), ‹x is knowledge› is inconsistent with ‹x isn't known through observation.› But, as we noted on page 24, observation cannot tell you whether one statement is consistent with some other statement. Thus, any body observational data is consistent with the assumption that empiricism is false.

This means that there cannot be strictly observational grounds for believing empiricism correct. Thus, so far as empiricism is correct, there are no grounds for believing it correct; and so far as there are such grounds, empiricism is false. Thus, the likelihood that empiricism is correct is inversely proportional to the degree of probability that the information at our disposal confers on it. And this means that, if it's a certainty that empiricism is correct, it's a certainty that it's false. Therefore, empiricism, if true, is false; therefore it's false.

The final sentence of a famous argument given by Bertrand Russell (1872–1970) ends with a sentence very similar to the last one. (This was deliberate.) The argument in question is to the effect that "naïve realism"—which is a specific, particularly extreme form of empiricism, and is therefore relevant in this context—is false:

> We all start from naïve realism, i.e., the doctrine that things are what they seem. We think that grass is green, that stones are hard, and that snow is cold. But physics assures us that the greenness of grass, the hardness of stones, and coldness of snow are not the greenness, hardness, and coldness that we know in our experience, but something very different. The observer, when he seems to himself to be observing a stone, is really, if physics is to be believed, observing the effects of the stone upon himself. Thus, science seems to be at war upon itself. When it most means to be objective, it finds itself plunged into subjectivity against its will. Naïve realism leads to physics; and physics, if true, shows that naïve realism is false. Therefore, naïve realism, if true, is false; therefore it's false.[73]

Though eloquently stated, this argument consists of spurious reasoning for a false conclusion. In observing the paper-weight on my desk, I'm not observing some effect of the stone upon myself: I'm observing the stone itself. To be sure, my observing the stone is *itself* an effect of some event involving the stone. (Light bounces off the stone and, in due course, strikes my retinas, precipitating various physiological and psychological responses, among them the aforementioned sense-perception.) But that doesn't mean that *what* I'm seeing, in having that sense-perception, is some effect that the stone had on me; and unless we're perpetually hallucinating, in which case we're never observing anything external, we very obviously do observe external objects.

This brings us to the second problem with Russell's argument. Russell says that the greenness, hardness, and coldness of daily observation are not identical with the counterparts in physics. This is false. Physics has a lot to say about greenness, hardness, and coldness that commonsense does not. But that's very different from saying that the coldness we feel, the greenness we see, etc., *aren't* the greenness, coldness, and hardness of physics. Physics tells us *what it is* for something to have the properties we know them to have through sight, touch, etc. When you grab and object and feel that it's cold, you don't feel or otherwise sense-perceive the micro-events *in virtue of which* it is cold. It's the job of the theoretical physicist to tell you about these microevents. This means that the coldness studied by the physicist *is* identical with the coldness that you feel. Physics tells us that many of our pre-theoretic beliefs as to what *that coldness is* are wrong; and in order to do that, it has to study the coldness that those pre-theoretic beliefs concern. So Russell's argument, despite Einstein's high regard for it, is a failure.[74]

4.7.2 The empiricism-unfriendliness of the concept of confirmation

Confirmationalism, the doctrine that a non-tautology is meaningful iff confirmable, is a form of empiricism. But we'll now see that the concept of confirmation is an incoherent one *unless* it's granted that there is nonempirical knowledge and, therefore, that confirmationalism is incoherent.

An argument due to Nelson Goodman (1954) makes this clear[75]:

> (GA[76]) Let's say that an object is "grue" if it's green and examined before Jan. 1, 2010, or it's blue and examined anytime thereafter. All green objects examined before Jan. 1, 2010, are grue. So, supposing that we've examined ten million emeralds before Jan. 1, 2010, and found them all to be green, we've also found them to be grue. Presumably, the fact that they've all been green warrants the inference that they'll be green after Jan. 1, 2010. But, so far as that data entitles us to infer that they'll be green, it also entitles us to infer that they'll be grue and, therefore, blue.
>
> This line of thought is easily extended to show that anything can confirm anything.[77] Let phi, psi, and chi be three properties such that (i) a thing is phi if it's examined by a human being who knows and therefore truly believes that, at that time, no human being can fly; (ii) a thing is psi if examined by a human who knows that, at that time, all human beings can fly; and (iii) a thing is chi if it's examined before Jan. 1, 2010, and known to have phi or it's examined after that time and is known to have psi. Since everything ever examined as of the present time (May 24, 2009) has had phi, it's also had chi. So given that, thus far, no human has been able to fly, we're no less entitled to infer that in 2015 they'll all be able to fly than we are entitled to infer that, at that same time, none of them will.

GA can be taken to show *either* that no inductive inference is better than any other *or* that, since some inductive inferences clearly are better than others, GA must involve an error of some kind. Supposing that the second interpretation is the right one, it's easy to identify the problem with GA. Contrary to what that argument tacitly assumes, induction does not have strictly observational basis. From a strictly *observational* standpoint, it's no less correct to describe an emerald examined in 2009 as "grue" than it is to describe it as "green." Given any body of data, there are different, but equally

observationally legitimate ways of describing it. This means that, if any inductions are better than any others, we must have legitimate but at least partly *non*observational grounds for believing that, when making inductive inferences, certain properties (e.g., *green*) are relevant and others (e.g., *grue*) are not. (These grounds are identified in Chapters 12 and 18.)

4.8 The brokenness of the concept of tautological truth

In this section we'll see that it is utterances of sentences, not sentences per se, that are tautologies. For reasons that will become clear, this entails that, contrary to what LP alleges, non-empirical truth cannot be identified with conventional truth.

Whether a given utterance is tautologous very much depends on the manner in which the person hearing that utterance learned the meanings of the expressions composing it. A story may help make this clear. You don't know to what length the word "yard" refers, and you ask your friend Smith to give you this information. In response, he points to some object L and says: "the length of *that* object is one yard." L is *in fact* three-feet long. But you can't tell this just from looking at it. You can tell roughly, but not exactly, how long L is. You don't bother to measure L. This all happens on Monday.

The next day, you see some object M. You measure it and find that its length is three feet. You tell your (still present) friend Smith that M is exactly three-feet long. You know, of course, that M's length is more or less comparable to L's—that neither length is, for example, ten times as great as the other. But you don't have precise knowledge of their comparative lengths; you don't know, for example, whether L's length is within six inches of M's. Because you have a passion for knowing the comparative lengths of objects, you find this upsetting, and you tell Smith that you wish you knew how L's length compared to M's. Smith says: "I don't know why you're upset. You've measured M and found that it's three-feet long. Since, as you know, L is a yard long, it's patently obvious what M's length is." You don't quite know what he means, and you tell him this. Somewhat irritated, he says:

(i) “there are three feet in a yard.”

Under these circumstances, (i) is *not* trivial and, therefore, is *not* a tautology. It *would* be tautologous if you had learned the meaning of the word “yard” by being told that “yards are lengths of three feet.” But this isn’t how you were told it. You were *shown* a yard-long object and told that the word “yard” refers to its length. Obviously that visual perception *did* apprise you of that object’s length. But the *way* it described that length to you was different from the *way* that this same information would be conveyed to you by an utterance of: “the lenght of *that* object is one yard.”

And given the information embodied in your visual perception’s of L and M, it wouldn’t be a trivial matter to know that the length described by the contents of your L-perceptions coincided with that described by your M-perceptions. Therefore, (i) would not, under those circumstances, express a tautology, at least not from your perspective. But it *would* express a tautology from the perspective of somebody to whom “yard” had initially been defined as “distance of three feet.”

So even though it’s standard practice among philosophers to describe *sentences* as “tautologies,” this practice embodies a serious confusion. A given sentence may or may not be tautology, depending on the manner in which the auditor learned the meanings of its constituent expressions and depending, therefore, on the information *on the basis of which* the auditor knows those meanings.

It might be thought that, so far as (i) is non-trivial to you, it’s only because you don’t *really* understand it. This isn’t true. To somebody who doesn’t speak Albanian, a sentence of Albanian isn’t trivial or non-trivial. It doesn’t mean anything to you and is, from your perspective, just so much noise and is no more “trivial” or “non-trivial” than the sound of wind chimes. But, in our story, you *do* know what is meant by (i). It’s not as though you’re hearing a sentence of Albanian (or, if you happen to speak Albanian, a sentence of some language that you don’t know). Therefore, it is only because you understand Smith’s utterance of (i) that it is nontrivial for you. Therefore, tautologousness, and non-tautologousness, are properties, not of sentences, but of the information on the basis of which auditors figure out the meanings of sentences. And it’s wrong to say that sentences *per se* are, or are

not, tautologies.[78]

This story illustrates some deeply important facts that are in the philosophy of language. First, one knows the meanings of expressions descriptively. It is through sight and hearing (and, possibly, other sensory modalities; e.g., touch) that you learn what words mean. Your perceptions apprise you of facts about the world by describing them to you—by apprising you of their colors, shapes, etc. Two very different descriptions can pick out some one thing (cf. "the third U.S. President" and "the President responsible for the Louisiana Purchase"). Therefore, the perceptually encoded descriptions *through* which one learns what two expressions mean may differ enormously, *even* if those expressions mean the same thing. A consequence is that what utterances tell you is as much a function of the information through which you learn their meanings as it is of the those meanings themselves. Thus, a given sentence may convey very different propositions to different people, all of whom know what it means, the reason being that those people access that meaning through different descriptions.

4.9 An alternative to the logical positivist conception of meaningfulness

The logical positivist's analysis of meaningfulness was a complete failure. I'd like to propose an alternative to it.

First of all, when asked to give examples of meaningless statements, the logical positivists tended to cite sentences that nobody ever uses; for example, "the nothing nothings"[1], "the all is one," "the absolute is perfect." This is deeply suspicious: a theory that only takes on straw men can't be much of a theory.

And LP can't even prevail against these straw men. For, contrary to what its advocates said, the problem with these so-called statements is *not* that they're incapable of empirical corroboration. Consider the sentence:

(i) "the universe is a perfect unity."

Much loved by many a freshman narco-intellectual, this is a meaningless sentence if ever there was one; and it is just the sort of sentence that logical

positivists had in mind.

But if it were said what exactly it means to describe something as a "perfect unity," (i) *would* be meaningful, as it would then be true or false. If, by a "perfect unity", one means an object that consists of events bearing a specified causal or logical relationship to one another, then (i) is either true or false, depending on the identity of that relationship. For example, if a "perfect unity" is an object such that, given any two nonsimultaneous events composing it, there is a possible causal process connecting the first of those two events with the second, then (i) is true.[1] (In contemporary physics, ‹x precedes y› is defined as: ‹there is a possible causal process, e.g. a light-signal, beginning with x and ending with y.›) On the other hand, if, by a "perfect unity," one means an object such that, if x and y are any two of its parts, the very idea of x's existing in the absence of y is an incoherent one, then (i) is meaningful—and false. (One can coherently imagine a universe in which Cheney exists but Biden does not.)

In any case, the term "perfect unity" is clearly intended to refer to some sort of causal or logical integratedness, and once that mode of integration is pinpointed, (i) speedily becomes a true or false claim.

"Statements" such as "the universe is a perfect unity," "the nothing nothings," and so on, aren't really statements at all. They're ***statement-forms***. (i) is obviously supposed to attribute some property to the universe. But since this property isn't identified, (i) contains an undefined term and therefore says nothing. Once that term is defined, a meaningful statement results.

It's true that (i) is neither tautologous nor confirmable. But that's a *consequence* of the real problem, viz. that "perfect unity" is undefined. (i) is comparable to "x is tall." The reason "x is tall" says nothing is that "x" is undefined—it hasn't been assigned a referent. And no sooner is a referent assigned to "x" than "x is tall" becomes meaningful. ("x is tall" comes to have the same meaning as "Bob Dole is tall" the moment Bob Dole is assigned to "x.")

Before "x" is assigned a referent, "x is tall" is neither confirmable nor tautologous. But that's only a symptom of the real problem, viz. that "x" doesn't have a referent. The same thing *mutatis mutandis* is true of (i).

Let's move onto the next phase of our argument. If S is a meaningful

sentence, there is some object x and some property phi such that S says that x has phi. In other words, any given sentence is equivalent to one that has the form: ‹x has phi.› Let us now say why this is so.

Any non-compound sentence (i.e., any sentence that doesn't consist of other sentences) either says of some individual that it has some property or it says that two more individuals are interrelated in a certain way. Thus, "Smith is tall" says of some individual (Smith) that he has a certain property (tallness), and "Bob loves Sally" says that one individual (Bob) bears a certain relation (that of loving) with respect to some other individual (Sally).

"Smith is tall," "Jerry snores," and all other non-relational, non-compound sentences obviously have the form ‹x has phi.› And, though it isn't obvious, the same is true of "Bob loves Sally," "Wilma detests Linda," and all other non-compound sentences that affirm the existence of relations between two or more objects. Let "R" be defined as follows: for any objects x and y, ‹ <x,y> has R› is true iff x loves y. (So for any objects x and y, the ordered pair <x,y> has R exactly if x loves y.) Thus, "Bob loves Sally" is equivalent with "<Bob, Sally> has R," which has the form ‹x has phi.›

A similar procedure can be performed on sentences (such as "Bob is standing in between Sally and Larry") that affirm the existence of relations involving three or more objects. Let "R*" be defined as follows: for any objects x, y, and z, ‹<x,y,z> has R*› is true iff x is standing in between y and z. Thus, "Bob is standing in between Sally and Larry" is equivalent with "<Bob, Sally, Larry> has R*," which has the form ‹x has phi.› Other non-compound relational sentences are to be dealt with similarly.

What about compound sentences? Not a problem. Let "K" be defined as follows: for any sentences S_1 and S_2, ‹ <S_1, S_2> has K› is true iff the state of affairs described by S_1 is a consequence (of some kind or other) of the state of affairs described by S_2. Thus, "Smith broke his leg because he fell out of the a tree" is equivalent with "<Smith broke his leg, Smith fell out of the a tree.> has K," which has the form ‹x has phi.› Other compound sentences are to be dealt with similarly.

Negative sentences are particularly easy to deal with. "Smith doesn't smoke" is equivalent with "the proposition that Smith smokes is false,"

which clearly has the form: ‹x has phi.› Other negative sentences are to be dealt with similarly.

There is only one kind of sentence that we haven't yet considered, namely, ***quantified generalizations.*** A "quantified generalization" is any statement that says how many members one class of objects has in common with some other class of objects. Examples are: (a) "some person smokes," (b) "no giraffes fly," and (c) "all mice read Tolstoy." (a) says that the class of people has at least one member in common with the class of smokers. (b) says that the class of giraffes has no members in common with the class of things that fly. And (c) says that the class of mice has no members in common with the class of things that don't read Tolstoy.

Bearing this in mind, let "E" be defined as follows: for any properties P and Q, ‹ <P, Q> has E› is true iff the class of things having P has a least one member with the class of things having Q. Thus, (a) is equivalent with: "<the property of being a person, the property of being a smoker>, has E" which obviously has the form ‹x has phi.› Other quantified generalizations are to be dealt with similarly.

We have thus established that any given sentence S is equivalent to some sentence having the form ‹x has phi.› Given this fact, there is an obvious answer to the question "what is it for a sentence to be meaningful?" A sentence is meaningful if it attributes some property to some object. A sentence S is meaningful if, for some object x and some property phi, S says that x has phi.[1] It's irrelevant whether it can be perceptually confirmed, let alone verified, that x has phi.

How could this theory be wrong? If a sentence attributes any property to any thing, it says something about something and is therefore meaningful. And if a sentence doesn't attribute any property to anything, it doesn't say anything about anything and is therefore meaningless.

5.0 The picture theory of meaning

One of the most interesting contentions put forth in the *Tractatus Logico-Philosophicus* (TLP) is the so-called "picture theory of meaning." In the TLP, Wittgenstein says that sentences are "pictures" of the facts they describe.

What does he mean? Maybe he means that sentences are picture-*like* in that they, like pictures, represent facts. But in that case what Wittgenstein is saying is completely and utterly trivial.

Thus, so far as what Wittgenstein is saying has substance, it isn't that sentences are *like* pictures of the facts they describe; and it must therefore be that they *are* such pictures.

But in that case, what Wittgenstein is saying is false. It is only relative to arbitrary conventions that "Smith punched Jones" describes the fact that Smith punched Jones. But it isn't relative to such conventions that a film or painting of Smith punching Jones describes that fact.

To be sure, there is a non-conventional *component* to sentential representation. Let P be the proposition meant by the sentence:

(1) "Given that Socrates was a philosopher, it follows from the fact that Socrates was bald that there has been at least one bald philosopher."

The fact that (1) means P is not itself a convention. It is a logical *consequence* of our semantic conventions (e.g., that "Socrates" refers to Socrates, etc.). But for that very reason, there is a conventional component to that fact. And since graphic resemblance is a *non*-conventional method of representation, it follows that, so far as (1)'s relation to the fact it describes is conventional, that relation is fundamentally *not* like the relation borne by a picture of an event to that event. Thus, the picture-theory is false if taken literally and it's empty if taken non-literally.

But maybe there's some way of interpreting that theory that we've overlooked. To see whether this is so, let's consider Wittgenstein's argument for it:

> At first sight a sentence—one set out on the printed page, for example—does not seem to be a picture of the reality with which it is concerned. But neither do written notes seem at first sight to be a picture of a piece of music, nor our phonetic notation (the alphabet) to be a picture of our speech. At yet these sign-languages prove to be pictures, even in the ordinary sense, of what they represent[79] . . .

> There is a general rule by means of which the musician can obtain the symphony from the score, and which makes it possible to derive the symphony from the groove on the gramophone record, and, using the first rule, to drive the score again. That is what constitutes the inner similarity between these things which seem to be constructed in such entirely different ways. And that rule is the law of projection which projects the symphony into the language of musical notation. It is the rule for translating this language into the language of gramophone records.

The idea seems to be that just as laws of projection coordinate the painting of the bowl of fruit with the bowl of fruit itself, so the ***semantic rules*** of a language coordinate its sentences with the realities they describe.[80]

But this analogy is a shallow one; and when it's scrutinized, it becomes even more clear than before how *un*like pictures sentences are.

What if, because of some change in the environment, snow turned black? The English language would not for that reason be impaired. In fact, the English language would, without itself having to change, give us the resources to describe this change. We could describe it by saying "snow is black." Thus, as far as the English language is concerned, snow can be any color. The semantic rules of English don't say that snow *is* white. They say that, *if* snow is white, one can express that fact by saying "snow is white." And those rules are to the effect that *if* snow is black, one can express *that* fact by saying "snow is black." So the semantic rules of English assigns sentences not to the *fact* that snow is white—for as far those semantic rules know, it isn't a fact that snow is white—but to the proposition that, when true, gives rise to that fact.[81] Thus, sentences depict facts *by way of* having propositions for their meanings. But this isn't how photographs work. A photograph doesn't go through the corresponding proposition. It goes straight to the fact (when there is one). No picture goes through a proposition. Pictures, unlike sentences, go straight to the facts, if any there be, that they represent. This shows how fundamentally unlike pictures sentences are.[82]

A related point is that pictures have structures that are radically different

from those of any sentences. Sentences are ***digital*** structures.[83] They have a unique decomposition into a finite number of discrete parts. ("The cat is on the mat" decomposes into "cat," "mat," etc.) Pictures aren't like this. A picture of a cat on the mat doesn't have one, minimal unit of significance corresponding to the cat, another to the mat, etc. The part of the picture corresponding to the cat may also contain a part corresponding to the cat's ear and to the cat's leg, etc.

The fact that sentences, unlike graphic representations, are digital structures is a consequence of the fact that the former, unlike the latter, have a conventional component. The reason for this is a subtle one. But it's worth stating, since it shows how deeply wrong the picture-theory is and since, in so doing, it reveals a lot about language.

Let D_1 be some random photograph of a person smiling. D_1 isn't a symbol of a language. But that could easily change. For this to happen, some convention would have to arise whereby it had a fixed a meaning— whereby it meant, for example, that people are sometimes are happy, and the same thing *mutatis mutandis* happens in connection with each of several other photographs. So for example, there is some photograph D_2 of a person who is crying, and some convention is created whereby D_2 means that people are sometimes unhappy; and so on. Let L be the language defined by the totality of these conventions.

Even though D_1 is an image, it isn't functioning as an image so far it's functioning as an expression of L. The fact that D_1 is a picture of a smiling person may obviously make it easier for people to remember that, in L, D_1 means that people are sometimes happy. But it won't be *what it is* for D_1 to bear that meaning, or any other, in L. The character "0" is an unfilled hole and can thus be taken as graphic representation of emptiness.

But so far as, "0" is such a representation, that isn't *what it is* for it to denote the integer preceding the number one. The same point *mutatis mutandis* holds of D_1.

Also, D_1 *doesn't* graphically represent the fact that people are sometimes happy; it graphically represents the tenuously related fact that, on some one

occasion, some one individual was happy, along with various other specific facts about that person's appearance that have nothing to do with anyone's being happy. So isn't by virtue of graphically representing the fact that people are sometimes happy that, when functioning as a sentence of L, D_1 describes that fact.[84]

So far as D_1 is a sentence of a language, what it actually picks out is irrelevant; the various nuances of the smiling gentleman's face are irrelevant. In general, its internal structure is irrelevant. Considered as an expression of L, it *has* no internal structure. (It is what philosophers of language call a **semantic primitive.** A semantic primitive," or primitive symbol," is one that doesn't consist of other symbols and that, so far as it is an expression of a language, thus has no internal structure.) For exactly similar reasons, each of the other photographs composing L is, when considered as an expression of L, devoid of internal structure and thus, in the most extreme way possible, *not* like a graphic representation.

We must make one more point before we can close the argument. Let N be a photograph of a bolt of lightning, and suppose that N is the L-translation of the English expression "it is not the case that." So if S is the L-translation of "grass is green," NS is the L-translation of "grass is not green." (NS is formed by putting N to the left of S.) Even though NS consists of pictures, it is not itself a picture. Putting two pictures together isn't one picture; it's just two pictures that are next to each other. For exactly similar reasons, if conventions were created whereby the sentences of L could be ***conjoined, disjoined***[85], or otherwise combined, the resulting compound sentences would not be pictures.

Let us take stock. Not a single one of the simple symbols belonging to L is a graphic representation of the fact that it depicts, and not a single one of the complex symbols belonging to L is a graphic representation of the fact depicted by any one of its components. In general, to the extent that a given thing is functioning as a linguistic expression, it is not functioning as a picture. Things that happen to be pictures cannot function as pictures so far as they are functioning as linguistic expressions or, therefore, as sentences. Thus, Wittgenstein's contention that sentences are pictures of the facts they describe is the antithesis of the truth. In addition to showing that

Wittgenstein's picture-theory is false, this shows that *any* conventional assignment of meaning to *any* collection of symbols—any *language,* in other words—necessarily yields symbols that have a unique decomposition into discrete parts. This is obviously, almost tautologically, true of compound symbols; and it's ***vacuously true***[86] of non-compound symbols, since no such symbol has any internal structure at all.

5.1 The picture theory of meaning (continued)

It's not entirely clear why Wittgenstein said that sentences are pictures. But it is clear that this contention of his is consistent with this empiricism.

Empiricism says that all knowledge is observation based. Thus, if you know it, you either (i) sense-perceived it or (ii) you inferred it from what you saw, provided that the inference rule you used is one that is known through sense-perception. (So far as knowledge is obtained with the help of inference rule whose legitimacy can't be authenticated by sense-perception, some knowledge is *not* perception-based.)

Our sense-perceptions give us pictures. Not all sensory modalities give us *visual* pictures, of course. Hearing gives us auditory pictures, touch gives us tactile pictures, etc. But perceptual representation is pictorial representation. (In what follows, when I say "see," I mean "see or hear or touch [etc.]."")

But much of what we know can't be embodied in images of any kind. (In this context, I'll use the word 'image' not just to still-images, but to moving pictures.) I know that:

(1) the moon is not made of cheese.

What would an image of this fact be? An image of a cheesy moon with a big X on it? No. The big X wouldn't be an *image* at all. Like the word "not," it would be a conventional sign of negation. Whereas a picture of a cheesy moon would indeed resemble a cheesy moon, a big X doesn't resemble the operation of negation. Nothing could physically resemble that operation, since it isn't something that could possibly be seen or otherwise sense-perceived.

Also, a picture of a cheesy moon corresponds to a lot of different

propositions.[87] Any such picture will also depict an object having a certain color, shape, etc. Since (1) doesn't anything about the moon's shape or color, it isn't identical with such an image. No proposition is identical with *any* image, since any image will contain information not contained in the image.

Thus, there are at least some cases where one's knowing of some fact doesn't consist in there being an image in one's mind of that fact. How is the strict empiricist to deal with this? First of all, it's hard to see how sense-perception, which gives us nothing but one image after another, could apprise us of truths that are incapable of being expressed in a strictly imagistic form. For argument's sake, let's concede to the empiricist that it's strictly through perception that I know that the Moon is made of XYZ. How are we to deal with my subsequent knowledge that the Moon is not made of cheese? Obviously that knowledge is *largely* based on my knowledge that it's made of XYZ. But it can't be *entirely* based on it. What my senses tell me, at most, is what the moon *is* made of, not what it isn't made of. So *some* kind of non-perceptual knowledge is involved in my making the leap from *the moon is made of XYZ* to *the moon is not made of cheese.*

Here is a related, if not quite coincident, argument. Even if image-resistant facts (e.g., those expressed by negative statements) are learned in a strictly perceptual manner, the mental states that mediate our knowledge of them are not themselves images. The information borne by those mental states must be encoded in some non-iconic form. This means that, at some point, pictorial information was converted into non-pictorial information. But if our post-perceptual mental states are to be knowledge, that conversion process must be a legitimate one. In other words, it can't, when given pictorial input x, yield some output y that is inconsistent with x. Moreover, we must *know* that the conversations being made are legitimate. For argument's sake, suppose that I don't *know* that, given my (let us assume) strictly perception-based knowledge that the moon is made of XYZ, it is *correct* to hold that the moon is not made of cheese. In that case, to the extent that my belief that it isn't made of cheese is based on my knowledge that it's made of XYZ, that belief isn't knowledge. If, on the basis of testimony given by a source of whose reliability I have knowledge, I believe P, I don't *know* that P. Uncorroborated testimony, though a helpful initial step on the road towards knowledge, is not itself enough for knowledge. For much the same

reason, if I don't know the rules that permit the derivation of non-perceptual beliefs from strictly perceptual ones, then, even if my post-perceptual beliefs are correct, they aren't knowledge.

But there couldn't possibly be any strictly perceptual way of knowing that those conversions were accurate. Those conversions, by supposition, turn pictures into non-pictures. So our knowledge of their existence, or (*a fortiori*) of their legitimacy, cannot itself be strictly pictorial. This is the real problem with empiricism. The rules that we use to make inferences *from* perceptual experience cannot *themselves* be learned strictly on the basis of sense-perception. This will be discussed at length in Chapters 12 and 13.

5.2 The picture theory of meaning (continued)

Interestingly, Wittgenstein made points at least vaguely like these in the TLP:

> In order to be ale to represent logical form, we should have to be able to station ourselves with sentences outside logic, that is to say outside the world. Sentences cannot represent logical form: it is mirrored in them. What finds its reflection in language, language cannot represent. What expresses *itself* in language, *we* cannot express by means of language. Propositions *show* the logical form of reality. They display it.
>
> Thus, if one proposition 'fa' shows that the object a occurs in its sense, two sentences 'fa' and 'ga' show that the same object is mentioned in both of them. If two sentences contradict one another, then their structure shows it; the same is true if one of them follows from the other. And so on. What *can* be shown, *cannot* be said.[90]

Wittgenstein seems to be saying that we cannot correctly describe the relationship holding between our words and the facts they describe. But, if correct, that point *itself* describes that relationship, calling into question its own coherence.[91]

In any case, contrary to what Wittgenstein says, we *can* identify the

logical forms of our own utterances. To identify the logical form of a statement is simply to make it clear what it means. We can do this. We do it all the time. We do it whenever we put the meanings of words into words. There are *some* qualifications to this, as we'll see in a moment, but none that redound to the credit of Wittgenstein's point.

Wittgenstein's assertion that "what *can* be shown, *cannot* be said" is obviously false. I tell you that I can do fifty push-ups. (I say "I can do fifty push-ups.") You don't believe me. So I show you that I can do fifty push-ups. (I do fifty push-ups in front of you.) And, as we just saw, that principle holds no less in connection with logical forms than in connection with a person's ability to do push-ups.

Echoing what we said earlier, although it isn't clear why Wittgenstein made these claims or what he meant by them, it is clear that they're consistent with his view that sentences are pictures of reality. A *picture* cannot picture itself. If P is a picture of a seagull, P can't contain a picture of itself, for the simple reason that nothing can be a proper part of itself. Of course, P might be a picture of a big seagull *and also* of some other, much littler, but otherwise identical seagull. But the big-seagull part of the picture isn't identical with the little-seagull part. The big part contains *two* seagull-images; the little part only contains one.

So i*f* sentences were pictures, then a given sentence S_1 couldn't be a picture of itself. But it doesn't follow that some other picture S_2 couldn't be a picture of S_1. Nor, therefore, does Wittgenstein's much stronger claim that *nothing*—no picture, no set of pictures, no sentence—could depict or otherwise represent the rules by which true statements are paired off with the facts they described.

Wittgenstein's claim that we cannot state the logical forms of sentences, which collapses into the brazenly false claim that we can never say what our words mean, has two roots. One of them is his just-discussed belief that sentences are pictures of the facts they describe. The other is his not yet discussed belief that, if it's assumed that we *can* say what our words mean, we have no way of dealing with paradoxes like the following. If somebody says:

(i) “what I’m saying is false,”

what that person is saying is true if it’s false and false if it’s true. Wittgenstein was keenly interested in this paradox during his pre-Tractarian years.[92] His reaction to it, it appears, was to hold that *any* attempt to articulate semantic rules would self-refer in the same the same paradox-engendering way as (i) and, therefore, that such rules cannot possibly be put into words.[93]

But this is not good reasoning. When I say, while pointing at the person exiting the limo, “that’s Mick Jagger,” I’m stating a semantic rule. There is some individual x such that I am saying (correctly, we may suppose) that it’s a semantic rule that “Mick Jagger” refers to x. The semantic rule I’m expressing doesn’t self-refer and isn’t otherwise defective.

As we’ll see in a moment, there are reasons to think that *some* semantic rules cannot be put into words, and Wittgenstein seems to have had at least a vague knowledge of some of them. But given only that *some* semantic rules can be put into words, it obviously doesn’t follow, contrary to what Wittgenstein seems to have inferred, that *no* semantic rules can be put into words.

Having spent pages dwelling on the shortcomings of the *Tractatus,* let’s end this section on a sunnier and more constructive note.

It is often said that truth is indefinable. Many people say this without meaning anything by it. Setting such people aside, those who say this seem to mean either (a) that it cannot be said what it is for a proposition to be true or (b) that no language can state all of the semantic rules that belong to it. (a) is false. (See Chapter 3.) But (b) is true.

(b) isn’t the absurd claim that no language can state *any* of its own semantic rules. Every time one uses an English sentence to define an English expression, one is expressing a semantic rule of English in English. But neither the English language, nor any other, can state *all* of its own semantic rules. Here’s why.

Given any meaningful expressions, there is a semantic rule r saying what s’s meaning is. (This is trivially true. A meaningful expression is one that has

a meaning and is therefore one such that some true proposition identifies that meaning; and any proposition that, like r, says what an expression means is *ipso facto* a semantic rule.) So, for example, supposing that x is Dick Cheney, there is a semantic rule of English to the effect that "Dick Cheney" refers to x. That rule is not *itself* a sentence. But it can obviously be expressed by a sentence (as we just saw). In general, semantic rules, though often capable of being expressed by sentences, are not themselves sentences. In light of this fact, suppose for argument's sake that, for each semantic rule of English, there is a sentence of English that expresses that rule. Let K be the class containing all and only sentences of English that correctly express actual semantic rules of English. Let SRE be the conjunction of all of K's members. SRE is a true and therefore meaningful sentence of English. SRE is also a member of K. After all, K contains every sentence that correctly says what is meant by at least one expression of English, and SRE obviously satisfies that requirement. But given that SRE is also a conjunction consisting of all of K's members, it follows that SRE is one of its own conjuncts. No conjunction can be one of its conjuncts. (The conjunction "snow is white and snow is white" is not one of its own conjuncts, since that sentence is a conjunction, whereas "snow is white" is not.) We're forced to reject this obvious truth *if* we grant the supposition that the English language can express each of its own semantic rules. Therefore, the English language cannot state all of its own semantic rules. Given any language L, what we just said about English is true of L. Thus, no language can express all of its own semantic rules.[94]

6.0 (ii) revisited: formal truth ≠ analytic truth

One of the main contentions of the TLP is that all ***entailment*** is *formal* entailment. One statement *entails* another if, supposing the first is true, the second couldn't possibly be false. So "Smith is a triangle" entails "Smith has more than one side." One statement, S_1, ***formally*** entails another sentence, S_2, if the statement "if S_1, then S_2" is a ***formal truth.*** A statement is a formal truth if every statement having the same form as it is true. Thus,

(1) "If Smith is in the barn, then it is not the case that it is not the case Smith is barn"

is a formal truth, since every statement of the form

(2) “if P, then it is not the case that it is not the case that P”

is true.

(1) is also an example of a formal entailment. It’s a formal entailment since it’s a formal truth that is also an entailment. (It’s an entailment since it’s to the effect that that one statement (*Smith is in the barn*) entails another (*it is not the case that it is not the case that Smith is in the barn*).

In the TLP, Wittgenstein asserted that all entailments are of this kind. He was aware that there are apparent counterexamples to this. For example:

(4) “Brown is a bachelor”

entails that

(5) “Brown is unmarried.”

But

(6) “if Brown is a bachelor, then Brown is unmarried”

isn’t formally true, since it has the same form as

(7) “if Brown is a bachelor, then Brown is a cupcake”

which isn’t true at all.

Wittgenstein deals with this by saying, very reasonably, that (4) is synonymous with:

(4_F) “Brown is unmarried and Brown is an adult and is male.”

Unlike (4), (4_F) *does* formally entail (5). According to Wittgenstein, all apparent counterexamples to his thesis that all entailment is formal entailment

can be dealt with similarly.

But the method used in connection with this particular counterexample fails in connection with others. The sentence:

(8) “Brown is a circle”

entails

(9) “Brown is a two-dimensional figure.”

(8) doesn’t *formally* entail (9). Wittgenstein must say that (8) is synonymous with something that *does* formally entail (9). If there is *any* sentence that is synonymous with (8) that formally entails (9), it’s:

(10) “Brown is a closed, planar, two-dimensional figure of uniform curvature.”

(10) does indeed formally entail (9). But (10) isn’t synonymous with (8). That’s why:

(11) “Brown is a circle iff Brown is a circle”

is trivial, and says nothing, whereas

(12) “Brown is a circle iff Brown is a closed, planar, two-dimensional figure of uniform curvature”

is non-trivial.

(12) doesn’t say anything about the *spatiotemporal* world. (12) is logically true; its truth is guaranteed by the structures of the concepts composing it. Unlike (11), (12) isn’t a tautology. Therefore, it’s a non-tautologous, non-empirical truth. Thus, there are non-formal entailments, and this entails that the Tractarian conception criterion of meaningfulness advocated false. (“Tractarian” is the adjective form of “the *Tractatus Logic-Philosophicus*.”)

7.0 Why the concept of a logically perfect

language is an incoherent one (This section is hard and should be skipped on a first reading.)

Let's say that a sentence is ***perspicuous*** iff its logical and grammatical forms coincide; and let's say that a language is ***logically perfect*** iff every sentence belonging to it is perspicuous.

Many non-perspicuous sentences belong to any given natural language (e.g., English, Swedish). One of Wittgenstein's objectives in the TLP is to identify the conditions that a language must meet if it is to be logically perfect. Wittgenstein takes it for granted that the concept of such a language is a coherent one. Influenced by the TLP, many early analytic philosophers longed for the day when logically perfect languages would replace natural languages.[95]

In any case, the concept of a logically perfect language is not a coherent one. It isn't possible for *everything* about a sentence's meaning to be reflected in its grammar. And if *per impossibile* there did exist a logically perfect language, it would be expressively inferior to English, Arabic, and every other natural language. In other words, there would be much that couldn't be said in it that could be said in any natural language.

More precisely, for each analytic truth capable of being expressed by a logically perfect language, there would be infinitely many that it could not express *and* that English or Spanish or any other natural language *could* express. At the same time, there would be no truth that could be expressed in a logically perfectly language that couldn't also be expressed in a logically imperfect language, such as English. Let us now discuss why this is so.

A sentence belonging to a logically perfect language is analytically true iff it is formally true. *Why is this?* A sentence is perspicuous only to the extent that its grammatical form makes it clear what it says and, therefore, makes it clear what it entails and what entails it. Bearing this in mind, let S be some arbitrary sentence. To the extent that there are false sentences, or true but non-analytic sentences, that have the same surface-structure as S, S's grammatical and logical forms don't coalesce and, consequently, S isn't perspicuous. So an analytically true sentence is perspicuous only if all other sentences having the same form are true. A sentence is formally true if, and

only if, any sentence having the same form is true. Thus, a sentence is perspicuous only if formally true. Therefore, a sentence belonging to a logically perfect language is analytic only if formally true.[96]

A consequence is that, for each analytic truth that a logically perfect language can express, there are infinitely many that it cannot express. Given any object x and any property phi, the sentence

(S_1) ‹if x has phi, then it is not the case that x does not have phi›

is perspicuous, the reason being that nothing having the same surface structure is false. But the superficially similar sentence

(S_2) "given any object x and any property phi, x has phi, then x does not have phi"

is *not* perspicuous, since it has the same surface structure as:

(S_3) "given no object x and no property phi, if x has phi, then x does not have phi,"

which is false. Formally true sentences are instances of informally true universal generalizations. So no sentence capable of expressing such generalizations is logically perfect.

This point has important and often overlooked consequences. Making a ***valid*** deductive inference involves recognizing an entailment. Your deductively inferring Q from P involves your recognizing that P entails Q. The only way to know that some formal entailment is valid is to know that some informal entailment is valid. You know that:

(S_4) "Jerry is in Richmond,"

entails that

(S_5) "it's not the case Jerry is not in Richmond"

And that's why, if you accept S_4, you also accept S_5.

But how do you know that S_4 entails S_5? Is it on the basis of your knowledge that all instances of S_1 are true? No. How could you possibly know that all of S_1's instances were true unless you could recognize the validity of specific inferences that it licenses (such as the inference from S_4 to S_5)? If you couldn't recognize that S_4 entailed S_5, then you obviously wouldn't have any idea why all of S_1's instances were correct.

This is not, at least not merely, a psychological point. It's a psychological corollary of an epistemological point, which, in its turn, is a corollary of a logical point. The reason why every instance of S_1 is true—the reason, in other words, why S_2 is true—is that and each of infinitely many specific inferences, of which the inference from S_4 entails S_5 is but a single instance, is valid. There are, quite literally, infinitely many informally valid inferences for each formally valid one. So Wittgenstein's allegation that all entailments are formal entailment is not feasible.

7.1 Hempel on the limits of strict empiricism

Carl Hempel (1905–1997) provided the following rigorous proof of the falsity of strict empiricism[97]:

> (HA[98]) If x's length is one unit, and y's length is √2 units, then there is no length L such that L goes an integral number of times into both x's length and y's length.
>
> Measurement is comparison with respect to some standard. To compare x's length with y's—in other words, to establish their comparative lengths—it is necessary to find some third body z that is taken as a standard[99]; and the relative lengths of x and y are determined by finding out how many z-length segments each of x and y can be divided into. (If one object can be divided into exactly twice as

many z-length segments as some other, then the first is twice as long as the second.) Supposing that x's length is one unit and y's length is √2 units, it follows that there is *no* body z such that both x and y can be divided, without remainder, into z-length segments.[100] It follows that, if an object's length is given by an irrational number, that fact cannot be known directly on the basis of measurement. It therefore follows that there is no strictly observation-based way to know that y's length equals x's multiplied by √2. In general, for any two objects x and y, there is no strictly observation-based way of establishing that x's length (or mass, etc.) is *incommensurable* with y's. (Two magnitudes M_1 and M_2 are incommensurable if there is no magnitude M_3 that goes an integral number of times into both M_1 and M_2.)

The branch of mathematics known as "calculus" is integral to modern physics. Calculus is the study of continuously changing quantities. In order to describe physical phenomena in a way that makes it possible to use the powerful techniques of the calculus to describe them, it must be assumed that they change continuously and, therefore, that the degree to which a given phenomenon has a given property may sometimes be given by an irrational number.[101] Thus, it must be assumed that, at certain junctures, the velocities, lengths, masses, etc., of objects are sometimes given by irrational numbers. But there cannot, as we've seen, be strictly observational grounds for believing that a given object's length is √2 meters (or that its mass is √2 lbs, etc.).[102] Since the calculus can't be applied to observable phenomena unless it's assumed that things' weights, velocities, etc., can at least sometimes assume values given by irrational numbers, it follows that modern physics integrally depends on an assumption for which

there cannot be possibly a strictly observational basis. This means that strict empiricism is inconsistent with the obvious fact that modern physics is a source of knowledge.

8.0 The sub-disciplines composing philosophy

The main branches of analytic philosophy are: the philosophy of mind, the philosophy of language, the theory of knowledge (also known as "epistemology"), philosophical logic, metaphysics, the philosophy of science, ethics, political philosophy, legal philosophy, the philosophy of religion, and formal logic.

It should be kept in mind that these sub-disciplines overlap a great deal. So, for example, the question "do we think in words?" belongs to the philosophy of mind *and* to the philosophy of language.

8.1 The philosophy of mind

This discipline studies the concepts in terms of which the mind is to be understood. Among the questions it tries to answer are:

Must one know a language in order to think? Or, on the contrary, is the ability to think a prerequisite to learning and operating with a language?

Given that knowing a language seems to enhance some kinds of thinking, how does it do so?

What are beliefs, and what is the difference between believing that Smith is tall and wondering whether he is tall?

How is perception related to thought?

Can perceptual content (i.e., what our eyes, ears, etc., tell us) be put into words? Or is there a fundamental difference between the kind of information that our sense-perceptions bear, on the one hand, and the kind of information that can be encoded in sentences, on the other?

To what extent is self-knowledge possible? What factors limit our ability to know ourselves? How is mind related to brain? Are they one? If not, what is the relationship between the two? Can there be unconscious mental activity?

8.2 The philosophy of language

This discipline studies the nature of linguistic meaning. Among the questions it tries to answer are:

What does it mean to say that "Smith" ***refers to*** Smith? What, in general, does it mean to say of an expression E that it picks out some object O?

How do the meanings of a sentence's parts relate to the meaning of the sentence as a whole?

Do expressions like "some person," "all people," and "no people" function in the same way as proper names (e.g., "Smith," "Jones")—that is, do they pick out objects? Or do they function in some other way? If so, what is that other way?

To what extent can the nature of linguistic meaning in general be understood in terms of the relationship that proper names bear to their referents (the things they refer to)?

How is it that statements about non-existent things can be meaningful?

How is the meaning of a sentence related to the thoughts of those who utter that sentence? Do the thought and the meaning coincide? Or is the relationship more indirect? If so, what exactly is that relationship?

How well does the grammatical structure of a sentence reveal what it actually says? Does grammar distort meaning or, on the contrary, is grammar a good guide to ***logical form***?

Are the ***semantic rules*** of a language (e.g., the rule that, in English, "snow" refers to a certain crystalline substance) known to speakers of that language? Or are such rules merely idealized descriptions of the behavior of those speakers?

Assuming, as some authors do, that there is an innately known language-like code in which we think, to what extent does that code resemble the languages (e.g., English, Spanish) that we *learn?*

How "transparent" is meaning? To what extent do users of a language know what sentences of that language mean?

What does it say of a sentence S that its *literal meaning* is P? What exactly is "literal" meaning? How is it different from communicated

meaning? Is literal meaning merely an idealized description of communicated meaning, or is it something else entirely?

8.3 Epistemology

This discipline studies the nature and extent of knowledge. Among the questions it tries to answer are:

What is knowledge? What separates those beliefs that are knowledge from those that are not?

What can be known and what cannot be know? (Can it be known what cannot be known? Or is it incoherent to give an affirmative answer to this question?)

Can we know about the future, the past, the possible but not actual, the impossible?

Can we know about the external world, or is knowledge limited to our own mental states? Can there be knowledge of things that are not in space or time (e.g., numbers) and, if so, how?

Are there any self-evident or self-justifying beliefs? Or must all justified beliefs be justified by beliefs other than themselves?

What is the structure of the *totality* of our knowledge? Are there some pieces of knowledge from which all the rest are derived or are all pieces of knowledge interdependent?

Is there a fundamental difference between knowledge of spatiotemporal fact (e.g., knowledge that there is a dog over there) and knowledge of purely conceptual truths (e.g., that there are laws only where there is government)? Or is the one kind of knowledge to be reduced to, or modeled on, the other?

8.4 Philosophical logic

This discipline studies *bearing-relations* holding among sentences and propositions (sentence-meanings). Among the questions it tries to answer are:

What is it for one statement to ***entail*** another? (P "entails" Q if there is no way that Q can be false if P is true.) Are there different kinds of entailment? If so, are some more central to reasoning than others?

Are inferences concerning the non-existent (e.g., “if Zeus is tall, then at least one god is tall”) to be modeled on inferences concerning the existent (e.g. , “if Bush is tall, then at least one president is tall”)? Are the same principles involved? Or is the non-existent logically *sui generis?*

To what extent can reasoning be “mechanized”? In other words, to what extent is it possible to produce rules that can be applied without any thought that will do the work of a rational being?

How are statements about what might have been, but is not, to be understood? Are they similar, logically, to statements about what is? Or do they have an altogether different logical form?

Are all statements either true or false? Or are some “indeterminate”—that is, is there a “gray zone”? And are there “degrees” of truth?

8.5 Metaphysics

This discipline studies the nature of possibility and necessity, of causal relations between objects. It also studies the nature of identity and the conditions that something must meet in order to exist. Among the questions it tries to answer are:

Under what circumstances are two distinct objects (e.g., my heart and my liver) both parts of some *one* thing?

What is it for an inanimate object to endure in time?

What is it for an animate object (e.g., a person) to endure in time?

Is there a sense in which fictional objects (e.g., Fred Flintstone) exist? Or is there no need to assume the existence of such things to account for the facts of experience?

What is it for something to be possible but not actual? What is it for something to be actual not necessary? What is it for something to be necessary?

Are necessity and possibility properties of objects (e.g., rocks, trees, people) or of statements? Are there things that are not in space or time?

Must things have causal properties in order to exist?

8.6 The philosophy of science

This discipline studies the logical structure of scientific endeavor and of its results. Among the questions it tries to answer are:

> What is the difference between statements that are scientific and those that are not? What are explanations? What is it to explain an event?

Is there a sharp distinction between theoretical and non-theoretical claims? Or, as some claim, are all statements (even basic ones; e.g., "that's a rock") "theory-infected"?

Given two rival theories, how is it to be determined which, if either, is the more accurate one? And supposing that one of them is the more accurate one, does it follow that it is the *better* one? In other words, is accuracy the only virtue a theory can have or, if not the only such virtue, then the most important one? Are theories to be evaluated (judged correct and, what may or may not be different, judged good) entirely in terms of their degree of agreement with the experimental data? Or are other factors (e.g., simplicity, comprehensiveness) involved?

What is the nature of measurement? Are there any reasons, other than reasons of convenience, for taking certain objects or events as standards? To use Hempel's (1952) example, is one wrong to take the Dalai Lama's heartbeat as a periodic process, or is it simply inconvenient to do so?

What is the nature of probability? What does it mean to say that there is a 50% chance that the coin will come up heads? Is probability just "a measure of ignorance," as Laplace (1749–1827) said? Or is it an objective fact about the world?

Do theoretical entities (e.g., protons, unconscious urges) exist in the same way as non-theoretical entities? Or are theoretical entities merely devices that we use to make sense of non-theoretical entities? Are statements about "protons" just abbreviated statements about meter readings and other macroscopic phenomena?

Is there a fundamental difference between explanations in the physical sciences and explanation in the psychological sciences?

Under what circumstances is a hypothesis (a tentative theory) to be rejected? Is a single disconfirmatory result enough? If not, what else is needed?

Must all theories be "deterministic"? (In other words, must they posit a rigid causal order?) (Einstein said "yes." Peirce (1839–1914) said "no.") Others, e.g., Ernest Nagel (1901–1985), say that the question is ill-formed, the reason being that whether a system is deterministic or not depends on how it is described. Determinism is a logical property of statements, in Nagel's view, not of the events they describe.[103] A consequence is that a given domain may be deterministic with respect to one method of describing it, but indeterministic with respect to some other method of describing it. Thus, the sub-atomic realm, Nagel says, is indeterministic with respect to the concepts in terms of which we describe the macroscopic realm; but it doesn't follow, Nagel plausibly alleges, that it is indeterministic *tout court*.)

Should science attempt to state how the world actually is? (Karl Popper says "yes.") Is that even possible? (Kant says "no.") Or should science confine itself to producing theories that are *consistent* with the data, while leaving it open whether those ***models*** are actually correct or not? (Bas van Fraassen says "yes."[104]) (A "model" is a description of a hypothetical structure that, *if* existent, *would* account for the relevant data.)

8.7 Ethics

This discipline studies the nature of good and bad, right and wrong. Among the questions it tries to answer are:

What is it for an act to be good and what is it for an act to be bad?

Are there absolute standards of goodness and badness, or do such standards vary from culture to culture?

Are there in fact such things as right and wrong?

Are any of our beliefs about the rightness and wrongness of things correct? Or are all our ethical beliefs illusions of some kind?

How are ethical statements (e.g., "killing is wrong") related to non-ethical, purely "descriptive" statements (e.g., "killing tends to undermine

social order")?

To what extent can one have ethical obligations towards oneself?

Does one have ethical obligations towards others, or should one be concerned only for oneself? To what extent, if any, is it in one's interest to act morally?

8.8 Political philosophy

This discipline studies the nature of law and government. It tries to identify the conditions under which laws and other political institutions are legitimate. Among the questions it tries to answer are:

What is a law?

What is the difference between a law and, for example, a gunman's threat?

What is a government? What is the difference between a government and, for example, the Mafia? How are legal rights related to ethical rights?

Can there be legal systems that are entirely evil, or must something embody at least a minimum of morality to qualify as a legal system?

Under what circumstances, if any, is one ethically entitled to break the law?

Under what circumstances, if any, does a government have the right to thwart the interests of its subjects?

What is the most just form of government?

Which kinds of freedoms ought a government to protect?

8.9 The philosophy of religion

This discipline studies the nature and existence of God and the conditions under which religious belief is justified. Among the questions it tries to answer are:

If there is a God, why do bad things happen?

Given that God, being invulnerable, cannot know what it is like to be

vulnerable, how can God know everything?

> If God knows everything, including what we will do, how can we have free will? If God is responsible for everything, how we can be justly punished for what we do?

> Does God have a gender? Does it make sense to say that God is a male as opposed to a female? Is there a God? If so, how is that to be established?

Is religious knowledge acquired in the same way as non-religious knowledge, or are different cognitive vehicles involved? And, once acquired, is religious knowledge (supposing such a thing to exist) to be *justified* in the same way as non-religious knowledge, or are different standards involved?

> What is the relationship between religion and morality? Can there be valid moral codes in a Godless world?

Can a genuinely religious person believe that God herself is bound by ethical principles? Or, in holding that God is so bound, is one undermining God's authority and, therefore, abjuring a religious outlook?

> Is acceptance of some kind of religion necessary for a meaningful life?

If there is an after-life of never-ending bliss, wouldn't we get bored? Does fulfillment involve adversity? Isn't struggle what gives life meaning?

8.10 Formal logic (a.k.a. mathematical logic, a.k.a. symbolic logic)

This discipline studies **formal truth.** The concept of "formal truth" is discussed in Chapter 7 and a precise definition of it is given in Chapter 18. But here's the basic idea.

S_2 formally follows from S_1 if the sentence ‹if S_1, then S_2› is formally true. A sentence is formally true if every sentence of the same ***form*** is true. A sentence has the same form as a given sentence if there is some ***open-sentence*** of which both sentences are instances. An open sentence is a sentence-like expression that contains a ***free variable*** and is thus neither true nor false. Synonyms of "open-sentence" are "statement-form" and "sentence-schema."[105] An open-sentence is formed by taking an actual sentence and

replacing one of the expressions in it with a variable. “Two is even” is an actual sentence. If the “two” is replaced with a variable, the result is ‹x is even›, which is an open sentence.

An ***instance*** of a sentence-form is what results when the variables in that sentence-form are replaced with constants. Thus, “two is even” and “five is even” are instances of ‹x is even.› To ***interpret*** an open-sentence is to replace the variables in it with constants, and an ***interpretation*** of an open-sentence is an assignment of constants to the variables in it. Consider the open-sentence ‹x has property phi.› An interpretation of that open-sentence is simply a proposal to the effect that the variables in it be replaced with constants. Thus, if I propose that the expressions “two” and “even” replace the first and second variables in that open-sentence, I am proposing an interpretation of it. Since the corresponding sentence (“two is even”) is correct, that interpretation ***validates*** that open-sentence. In general, an interpretation of an open-sentence validates it if the corresponding sentence is correct.

Not every interpretation of ‹x has property phi› validates it. For example, the interpretation of it that generates “two is odd” fails to do so.

If a given open-sentence is validated by every interpretation of it, then each instance of it is formally correct. This coincides with our earlier definition of “formally correct.” If every interpretation of a given open-sentence is correct, that open-sentence is said to be “true under all its interpretations.” It must be kept in mind that this is just a figure of speech, since open-sentences are not, in fact, true.

Statement-forms fall into three categories: (i) those whose instances are sometimes, but not always, correct (e.g., ‹x is even›); (ii) those whose instances are always false (e.g., ‹x is even but not divisible by two ›): and (iii) those whose instances are always correct (e.g., ‹x is identical with x›.) An open-sentence falls into (i), (ii), or (iii) depending on whether it is (i*) true under some, but not all, of its interpretations; (ii*) true under none of its interpretations; or (iii*) true under all of its interpretations.

Formal logic tries to ***formalize informal analytic truth,*** so far as that’s possible to do so, and to say when it isn’t possible, so far as it isn’t. An analytic truth is one whose negation is incoherent, and an analytic truth is informal if it has the same form as some false statement. Thus, “triangles

have three sides" is analytic, since "triangles don't have three sides is incoherent," and it's informal, since it has the same form as "squares have three sides." To formalize an informal analytic truth T is to identify an open-sentence S such that every instance of S is true and such that one of S's instances is equivalent with T.

Consider the sentence:

(1) Bill is self-identical.

(1) is an analytic truth, since its negation is

(2) Bill is not self-identical,which is incoherent.

But (1) isn't formally correct, since it has the same form as:

(3) Bill is green,

which is false, given that Bill is a non-green person. (1) and (3) are both instances of the form sentence-form:

(4) Bill has phi,

Since some of (4)'s instances are false, (1), though analytically true, is not formally so. But (1) is *equivalent* with a formal truth, namely:

(5) Bill is identical with Bill.

(5) is an instance of the form:

(6) x is identical with x.

The reason that (5) is formally true is that it's an instance of (6) and no instances of (6) are false. We just formalized an informal analytic truth and, therefore, did on a very small scale what mathematical logicians do on a very

big scale.

What we believe to be *bona fide* statements sometimes turn out to be statement-forms; and statement-forms that we believe to have only true instances sometimes turn out to have false ones. [106] Both of these deeply important facts first became apparent when, in the middle of the 19th century, Euclid's ***axiomatization*** of geometry was re-examined. Euclid showed that a great many geometrical truths follow from a small set of assumptions. These assumptions were:

(1) Any two points can be connected by a straight line-segment.

(2) Any line-segment is a part of some line.

(3) Given any point and given any line-segment starting from that point, there is a circle whose radius is the length of that line-segment.

(4) All right-angles are equal to each other.

(5) Given a line L_1 and a point P not on L_1, there is exactly one line L_2 that passes through P and doesn't intersect with L_1

(5) is known as the "parallel postulate."

We'll use the expression "(1)-(5)" to refer to the *conjunction* of (1) and (2) and (3), etc. Thus, (1)-(5) is a *single* open-sentence, and "(1)-(5)" is thus a singular, not a plural, noun. One would think that (1)-(5) is correct. But this turned out not to be so. It turned out that (1)-(5) is a statement-*form,* not a statement proper, and that (1)-(5) therefore isn't true *or* false. It also turned out that some of its instances are false. Let us now describe one such instance.

Let S be some sphere. Given an arbitrary point on S's surface, there is a path leading from that point back to that same point that cuts S into two symmetrical halves. If by a "line" we mean such a path, and we make the corresponding changes to the otherwise unchanged meanings of (1)-(4), the propositions thereby assigned to (1)-(4) are true, but the proposition assigned to (5) is false. For, if "space" and "line" are so defined, a line has zero parallels, as opposed to one.

According to many, this shows that the parallel postulate isn't true of

every possible space. This isn't what it shows. The parallel postulate isn't true or false of anything. It's a statement-form, not a statement, and statement-forms aren't true or false of anything. The right conclusion to draw is that there are possible spaces that are (partly) described by sentences that are negations of *instances* of the parallel postulate.

If "space" and "line" are defined in the conventional, Euclidean way, the sum of the interior angles of a triangle is 180. But if "space" and "line" are defined in the way just proposed, that sum may be anything greater than 180° and less than 360°. The larger the triangle, the greater the sum.

In (1)-(5), the words "line" and 'space' are functioning as variables, not as constants. This is an immediate consequence of the just-seen fact that, depending on what specific meanings are assigned to those words, (1)-(5) may come out either true or false. "But a 'space' isn't the surface of a sphere," one might protest. "And a 'line' isn't a pathway of the sort just described. So all you've shown is that by *mis*interpreting (1)-(5), you can generate some interesting results. But that means that you haven't really shown anything." Not true. In saying that "space" and "line" don't have these non-Euclidean meanings, one is making assumptions as to the nature of space that it is the very purpose of (1)-(5) to *establish.* In presupposing that ‹x is a space› entails that x isn't the surface of a sphere, one is in effect presupposing that triangles have interior angles adding up to 180°—one is, indeed, presupposing a great many of the principles that (1)-(5) are supposed to establish. So one cannot, without invalidating one's attempt to ground geometry in (1)-(5), assume that "line" and "space" are not to be defined in this way.

In (1)-(5), the words "line" and "space" are functioning as variables, not as constants. We've seen that (1)-(5) isn't true for all values of those variables. That is, some instances of that open-sentence are false. That is, that open-sentence isn't true under all its interpretations. Formal logicians aspire to identify open-sentences that *are* true under all their interpretations, since it is only to the extent that they can do this that they can formalize analytic truth, which is their main objective. In the course of this search, they inevitably come across many open-sentences, such as (1)-(5), that they had hoped were true under all their interpretations but turned out not to be. Given an open-sentence S of this kind, they try to say, as precisely as possible, what

it is that all those interpretations of S that validate it have in common with one another that they *don't* have in common with any interpretation that fails to validate S. In other words, they try to come up with a *general* characterization of "truth under S."

There are some classes of true statements that one would *expect* to be formalizable but turn out not to be. The class of arithmetical statements ("1 + 2 = 3," "2 × 9 = 18," etc.) is an example. In other words, arithmetical truth cannot be formalized.

First of all, arithmetical statements, as they are ordinarily expressed, are not formally true. "2 + 2 = 4" has the same form as "2 + 2 = 5." They both have the form "x + x = y." Since "2 + 2 = 5" is false, and has the same form as "2 + 2 = 4," the latter, though true, isn't formally so.

Formalizing arithmetic would involve finding some open-sentence S such that, for some interpretation of S, every true arithmetical statement is a formal consequence of that interpretation and such that no false arithmetical statement is such a consequence.

It turned out that this is not possible. This means that any formal characterization of arithmetic is either ***inconsistent*** (i.e., it entails a contradiction), or ***incomplete*** *(i.e., there is some arithmetical truth that is not a consequence of it)*. (The reasons for this are outlined in Chapter 7.) Given a body of truths that might appear to be capable of being formalized, mathematical logicians wish prove whether or not it is so; and, supposing that it can be formalized, they wish to find a model for it.

Chapter 2

Properties and Non-Spatiotemporal Existence

1.0 Why we must grant that there are non-spatiotemporal entities

Not everything that exists is in space-time. Everything that is in space-time is an ***instance*** of at least one ***property,*** and properties are not themselves in

space-time. I like to play tennis. I thus have the property of liking to play tennis and am thus an instance of that property. I am obviously in space-time, but that property is not. No matter how thoroughly one searches the universe, one will never encounter the property of liking to play tennis. One will, at most, encounter its instances.

Many hold that, although property-instances exist, properties *per se* either don't exist or exist but only because they're identical with spatiotemporal entities of some kind (e.g., aggregates of their own instances, ideas in people's minds). In this chapter, we'll discuss why this seemingly reasonable viewpoint cannot be sustained. We'll also see why the negation of this viewpoint is, contrary to initial appearances, utterly consonant with commonsense.

Properties themselves have properties. The property of being a square circle has the property of being uninstantiated. Properties of properties are ***higher-order properties.*** Properties that are not higher-order properties are ***first-order*** properties. The instances of higher-properties, being properties themselves, are not in space-time. The instances of first-order properties are in space-time.

Everything that exists is either a property or is in space-time. Non-spatiotemporal entities that *seem* not to be properties always turn out to be so. For example, ***propositions*** are not spatiotemporal and propositions do not initially seem to be identical with properties. (Propositions are sentence-meanings.) But propositions turn out to be properties. (We'll see why in Chapter 3.) ***Sets*** are not spatiotemporal, and sets do not initially appear to be identical with properties. But they too turn out to be properties. (But, contrary to what many a philosopher has said, properties are *not* identical with the sets that contain their own instances. The property of being a bird is not identical with the set of all birds. It's identical with a very different set, as we'll see at the end of the present chapter.)

The doctrine that spatiotemporal things are instances of non-spatiotemporal things is known as "Platonism." This is because Plato was the first person to advocate that doctrine. Sometimes Platonism is known as the *theory of universals,* and we will sometimes refer to it that way. The doctrine that properties don't exist, or do exist but only because they are identical with

spatiotemporal entities, is sometimes referred to as “nominalism.” But we’ll usually refer to it as “anti-Platonism.”

Incidentally, nominalism is so called because some anti-Platonists held that universals are identical with word. (“Nomen” being the Latin word for “name” or “word.”) In Section 3.0, we’ll see that this position, in addition to being false, is incoherent, the reason being that words are themselves universals.[49]

1.1 Why we must grant that there are non-spatiotemporal entities (continued)

Given that:

(1) Sally and Fred are both intelligent

it follows that there is something that they have in common; that is, that

(2) there exists some property P such that Fred is an instance of P and Sally is an instance of P.

Instances of intelligence are in space-time. Einstein was in space-time; so was Keynes. But intelligence itself is not in space-time. “But mightn’t the property of intelligence be nothing other than some sort of aggregate of all its instances?,” one might ask. “Mightn’t there be some way of spatio-temporalizing that entity?”

No. Consider the property of having a perfectly triangular shape. Nothing has this property. Are we to say that it doesn’t exist? Were we to do so, we’d be saying that, since there are no instances of the property of triangularity, that property doesn’t exist. Which is the same as saying there exists a property, viz. that of triangularity, such that, because nothing has it, it doesn’t exist. Which is absurd.

Also, a consequence of identifying properties with “aggregates” of their instances is that properties that clearly aren’t identical with each other must be seen as being identical with each other. (In this context, the word

"aggregate" can refer to any spatiotemporal entity that is composed, in *any* sense of the word "compose," of other spatiotemporal entities.) There are no perfectly spherical objects; nor, as we just pointed out, are there any perfectly triangular objects. So, if we identify properties with aggregates of their instances, the aggregate of all the instances of the one property is identical with the aggregate of all the instances of the other, the reason being that each aggregate is a perfect nullity.

A story will clarify these points. Max is the world's greatest harpsichord player. He's also the only redhaired person who ever has, or ever will, set foot in house X. The property *red-haired person who sets foot in X* is co-extensive with the property *world's greatest harpsichord player.* Given that, for this reason, there is *no* aggregate of things having either property, it follows that no aggregate of things having the one property is different in any respect from any aggregate having the other property. This means that, anything x that is an aggregate of the one kind will be identical with anything y that is an aggregate of the other. Thus, a consequence of identifying properties with aggregates of their instances is that there cannot be distinct, uninstantiated properties. But that's absurd. The property of being a perfect husband is obviously distinct from the property of being a perfect soldier, even though neither property is instantiated.

1.2 Logic property-based, not set-based

Consider the clearly correct statement that:

(1) squareness is a form of rectangularity.

On the face of it, (1) seems to make a statement about the property of squareness. So, if (1) 's appearance is to be trusted, then there exists a property (viz. that of squareness) that has another property (viz. that of being a form of rectangularity).Many philosophers[50] have tried to show that, despite its appearance, (1) doesn't really talk about properties *per se* and that it instead talks about their instances. They've proposed that the real meaning of (1) is:

(2) everything that is square is a rectangle.

And they've proposed that other statements that, like (1), seem to refer to properties be treated similarly. (So "green is a color," they say really just means that everything green has a color.)

But (1) doesn't mean the same thing as (2). If there were exactly two square things, x and y, and x and y happened to be green, the statement:

(3) everything that is square is green

would be true, but

(4) squareness is a form of greenness

would be false (and so would other statements of its ilk; e.g., "squareness is a color").

Of course, the statement

(5) everything that is square is *for that very reason* a rectangle

is true, whereas

(6) everything that is square is *for that very reason* green

is false, even if the only two square things happen to be green.

What (5) says is that:

(7) a thing's being a square is responsible for its being a rectangle,

which is obviously correct; and (6) says that:

(8) a thing's being a square is responsible for its being green,

which is obviously false. But (6) collapses into (1). (6) says that, to the extent that everything square is rectangular, it's because the property of squareness bears a certain relation (that of being a form of) to the relation of rectangularity. The relationship between a thing's being a square and its being a rectangle is, (6) rightly says, to be understood in terms of a relationship holding between squareness and rectangularity. But the relationship between a thing's being square and its being *green* is not to be understood in such terms.

A thing is a rectangle *in virtue of* its being a square, but a thing isn't

green not *in virtue of* its being a square even if it happens to be a square. Whenever one encounters an expression like "in virtue of," "*ipso facto* ," "to the extent that," or "for that very reason," one knows that one is dealing with a statement that is primarily about properties and only secondarily about objects. If I say that "copper things are *ipso facto* good conductors of electricity," I'm saying that the relationship between a given thing's being made of copper and its being a good conductor of electricity isn't fortuitous and is grounded in the natures of those two properties. In other words, to understand why an instance of the first property is an instance of the second, one looks to the natures of those properties. The connection goes from property instance (instance of being made of copper) to property (property of being made of copper) to property (property of being a good conductor of electricity) to property-instance (instance of being a good conductor of electricity). And the reason why "copper things are *ipso facto* in JMK's living room" is false is that, even though (so we will suppose for argument's sake) every copper object happens to be in my living room, the connection between x's being made of copper and x's being in my living room isn't routed in that way through those properties.

All *principles* connections are relationships between properties. This means that ***nomic***[51] connections (e.g., the connection between a thing's being made of copper and its being a good conductor of electricity) and also all logical connections (e.g., the connections between a thing's having knowledge and its having a mind) are expressed by statements whose subjects are properties. The correct statement that:

(9) For any x, the statement *x has knowledge* entails *x has a mind*

is another way of saying that

(10) something that has knowledge *for that very reason* has a mind of some kind.

And (10) is another way of saying that:

(10) the property of having knowledge is a form of the property of having a mind.

In conclusion, unless it's conceded that properties exist, there is no way

to account for the fact that there are *principled* relationships of any kind. And in fact those who deny that properties exist tend *not* to grant that there exist any such relationships. Those who deny the existence of properties tend to be ***empiricists.*** Since empiricists cannot grant the existence of things, such as properties, that cannot be seen or otherwise known (more or less) directly on the basis of observation, they end up *denying* that anything has any bearing on anything. (See Chapter 18.)

1.3 Logic property-based, not set-based (continued)

Let S_1 be the set containing the smallest number greater than zero that is divisible by two, and let S_2 be Fred's favorite number, which happens to be the number two. S_1 and S_2 are the same set, and the statement:

(1) x is a member of S_1

is therefore identical with the statement

(2) x is a member of the set that contains the number two and doesn't contain anything else;

and both statements are identical with

(3) x is a member of S2.

Thus, anything entailed by any one of those statements is entailed by each of the others, and anything that entails any one of those statements entails each of the others. Thus, (1) entails:

(4) x is less than twenty,

and so do (2) and (3). And

(5) x is a whole number greater than one but less than three

entails (1), and it also entails each of (2) and (3).

But the statement:

(6) x has the property of being Fred's favorite number

entails

(7) x has the property that Fred is aware of its existence,

but (1) doesn't entail this, and neither do (2) or (3). And

(8) x has the property that it is a number that is preferred by Fred to every other number that Fred has considered

entails (6), but it doesn't entail (1) or (2) or (3). It follows that properties cannot be identified with sets of their instances without mutilating the most basic principles of logic.

"But surely that's wrong," one will say. "(1) is the statement that x is a member of the set that contains the smallest number divisible by two (and nothing else)," and (2) is the statement that x is a member of the set containing Fred's favorite number (and nothing else). So I don't see how (1) and (2) could entail, and be entailed by, the same things; nor do I see how (1) and (6) could entail, or be entailed by, different things."

If the expression "the set containing Fred's favorite number" is taken in this way—in other words, if it's taken in such a way that (1) and (6) are equivalent, whereas (1) and (2) are not—then the word "set" is going proxy for the word "property." If by "the set containing Fred's favorite number" you mean, *not* "the set containing the number two," but instead "*any* set, whatever its membership, that comprises Fred's favorite number," then what you mean by "the set containing Fred's favorite number" is: "any set that is generated by the *property* of being Fred's favorite number." But in that case, when you make statements about "sets," you're really talking about properties.

1.4 The axiom of extensionality

It may be obvious at this point that properties aren't identical with sets of their instances. But this now obvious point is easily extended to reveal the futility of attempts to show that properties, if existent, are spatiotemporal. But in order to make the needed extensions, we must take a very quick detour through a bit of set theory.

Two very different properties can be ***co-extensive.*** Two properties are co-extensive if any instance of the one is an instance of the other. But different *sets* cannot have the same members. This is known as the ***axiom of***

extensionality. Since different properties can generate the same set, properties aren't identical with sets of their instances. We'll revisit this point later in this chapter. So if S_1 is a set and S_2 is a set, then S_1 is identical with S_2 if and only if nothing is a member of the one that isn't also a member of the other. A set whose members are x, y, and z can't fail to have those members any more than it can fail to be itself—indeed, for it to fail to have those members would be for it to be fail to be itself. Thus, sets are *individuated* by their memberships. This principle is sometimes known as the axiom of extensionality.

A corollary of this axiom is that a set's membership is frozen—frozen, indeed, in two senses. First, it's temporally frozen. Sets don't gain members and they don't lose them. Second, it's *modally* frozen. If a set's actual members are $x_1...x_n$, it couldn't possibly exist without having exactly those members.

This may at first seem patently false. "Surely the set of people changes from year to year; people are born; people die." This is the wrong view. The right view is that some one *property*—that of being a person—generates *different* sets at different times. So really there is no such thing as *the* set of people. There is the set of people at 3:00 P.M. April 28, 2009, the set of people at 3:01 P.M. April 28, 2009, etc.

And if S is the set of people—in other words, if it's the set that contains you, me, Barack Obama, etc.—S couldn't possibly exist in a universe in which one of us three didn't exist. Of course, there could be *people* in a world where one of us didn't exist. And, in that world, the *property* of being a person would generate *a* set. But that set wouldn't be S. It would be some other set.

"But isn't the axiom of extensionality just a matter of convention? Couldn't we stipulate that sets *could* lose and gain members?" If we used the word "set" in this way, it would become a proxy for the word "property." To say that the "set" of humans changes every time someone is born or dies is to say that, at any given time, this set's membership is determined by what it is that, at that time, has the *property* of being human. So, while there's nothing wrong with using the word "set" in this way, it isn't an option for somebody who wishes to replace properties with sets.

Having taken care of these technical preliminaries, let us move on to matters of substance. If some piece of copper happens, at a given time, to be conducting electricity, it's because of what is going on right then and there. The fact that there exists some other piece of copper in some remote galaxy is irrelevant. But if S is the set of all pieces of copper, the existence of that other piece of copper *is* relevant to S's existence. Sets are individuated by their memberships. (The set consisting of Smith, Larry, and Sally ceases to exist if Smith does.) So if, for some reason, the universe hadn't begotten that other piece of copper, then S wouldn't exist. But what difference would that make to the fact that *this* piece of copper is conducting electricity? None.[52]

Any given thing belongs to many different sets. What those sets are is only as good an indicator of how that thing is as it is an indicator of what properties that thing has. If S is the set of people who actually exist, then S doesn't exist in a world W that where Dick Cheney doesn't exist but is otherwise exactly like our world. But what difference does that make to my life-path in W? None. So far as you can learn anything about me from the fact that I belong to the set of humans, it's because, in learning this, you are learning that I have the property of being human.

Everything just said about sets can be said of *any* entity that is constituted by property-instances. Suppose, for argument's sake, that there is a "scattered object" consisting of all and only those human beings in existence. (This is a popular view among philosophers.[53]) Just replace each occurrence of "set" in the argument just given with "scattered object consisting of all and only human beings," and the resulting argument shows that the property of being a person isn't identical with that set. (Also, we'll see in the next section that, setting aside properties that have only one instance, there are *no* spatiotemporal entities that can be identified with a given property's instances. There are no "scattered objects.")

2.0 Arguments against Platonism

It is seldom denied that there are instances of properties, but it has often been denied that there are non-spatiotemporal entities. But it's hard to see (i) how there could be instances of properties without there being properties themselves, and it is also hard to see (ii) how properties could be given

spatiotemporal coordinates.

Opponents of Platonism respond either by saying that (i) involves some sort of logical or linguistic trickery or by saying that properties *can* be identified with spatiotemporal entities. Let us now consider how anti-Platonists defend their view.

2.1. Argument #1 against Platonism

Not every word in a sentence corresponds to some entity in the world. If I say "for the sake of all that is good, make sure you give at least some money to charity" the word "sake" doesn't denote anything—there aren't such things as sakes, after all—even though, being a noun, the word "sake" has the same syntactic function as words that do have denotations.

We can develop this point by taking a brief tour through linguistic theory. It seems fairly clear that not all expressions *refer* to things. Consider words like "of," "and," "the," "nonetheless." It's hard to believe that these words are comparable to "Smith," "Jones," and "cow." The medievals referred to expressions like "Smith" and "Jones" as "categorematic" expressions, and to expressions like "of" and "the" as "syncategorematic" expressions. A syncategorematic expression is one that doesn't refer to anything and whose linguistic function is strictly syntactical. Contemporary linguists make a similar distinction: they distinguish between "lexical" and "non-lexical" items: the former include "Smith," "Jones," "red," and "cow"; the latter include "the," "of," and "nonetheless." A lexical item is one that has a referent; a non-lexical item is one that does not, and whose purpose is strictly syntactic. (In other words, the purpose of a non-lexical items is simply to facilitate the unification of lexical items to form larger lexical items.)

Even though there are many expressions that appear to denote properties, they are really syncategorematic non-lexical expressions. Words like "red" or "the property of redness" don't refer to anything. This is not to say that they are not meaningful. "Of" is meaningful, even though it doesn't denote anything, and the same is true of "the property of redness."[54]

This argument evaluated: Expressions like "red" and "the property of redness" have very different logical properties from expressions that are

known to be syncategorematic. If I know that:

(1) x is red

is true, then I can logically infer

(2) There exists some color or other that x has.

This inference involves affirming the existence of the thing meant by the term "red." Many inferences have this structure. Given:

(3) Fred is smart but Larry is not,

I can infer that

(4) there exists some property that Fred has that Larry lacks.

Words like "of" don't behave in a comparable manner. Consider:

(4) There is a picture of Paul on the mantelpiece; nonetheless, I don't think that the owner of the mantelpiece likes Paul, since the picture is an extremely unflattering one.

I can make many inferences from (4), but not a single one involves affirming the existence of something meant by "of" or "nonetheless."

So it is not easy to make a case that words like "smart" and "red"—words that denote properties—lack referents and are really to be thought of on the model of purely syntactic terms like "of."

There is another problem with Argument #1.Words are themselves universals. The word "red" isn't identical with this or that ink mark or burst of noise. Given any inscription or burst of noise—given *any* spatiotemporal entity—that word can exist even if that spatiotemporal entity is destroyed. When you erase an inscription of the word "red," you don't erase *the word itself.* Inscriptions of words, and spoken utterances of them, are *instances* of words, not words *per se.* Words themselves are *properties* of such physical entities. We'll pick this up later.

2.2 Argument #2

Although properties exist, they are identical with spatiotemporal entities.

> The property of being a rock is identical with the *set* of rocks. And the set of rocks is a spatiotemporal entity. That set, after all, is simply a very spread out object that consists of rocks.

This argument evaluated: (The first two paragraphs of this section repeat many of the points made in Sections 1.0–1.5.) This answer is no good. Sets are abstract objects. They don't have space-time coordinates. The *set* of rocks is not some big rock and it isn't some big object consisting of rocks separated by stretches of empty space. If all the rocks in the world move to some new place, and none are destroyed, the set of rocks hasn't changed at all: it is the same as it was before. But, under those circumstances, any object consisting of all the rocks in the world *has* changed: it has acquired a new location (or, in any case, it has undergone an internal structural change).

A set is a mathematical entity; it has no location and is defined entirely by its formal properties. *Where* its members are, and what they are doing, is irrelevant (so long as its members don't cease to exist). But any object that is actually *composed* of all the rocks in the world *is* affected by the movements of their members, and thus cannot be identified with sets. (Of course, it is unclear whether there is such an object—an issue we'll revisit shortly—but what is clear is that *if* there is, *then* it is affected by movements of the kind just described.) Thus, sets aren't spatiotemporal entities. Consequently, they are *themselves* Platonic entities. So one is conceding the truth of Platonism, not undermining it, by identifying properties with sets.

Aware that, for the reasons just given, sets cannot be identified with sets, philosophers of an anti-Platonist bent have sometimes chosen to identify properties with what are referred to by contemporary philosophers as *scattered objects.*

What is a scattered object? Ordinary objects are geographically and dynamically cohesive. Consider your back-pack. If you move one of the straps ten feet across the room, you have moved the whole bag across the room (unless, of course, you've damaged the bag in some way). And even though (at the microscopic level) the bag consists mostly of empty space, all of its parts occupy one, fairly compact, relatively continuous region of space-time.

Now consider all the rocks in the universe. Those rocks clearly *don't*

constitute a dynamically or geographically integrated unit. Supposing that those rocks collectively constitute any object at all, the object they constitute is a "scattered object." In general, a scattered object is one that lacks the cohesiveness and dynamic integrity of ordinary objects. According to some anti-Platonists, properties are scattered objects. Is this correct?

No. There are no scattered objects. The very concept of such a thing is an incoherent one. In Chapter 16, we'll see why. Here's an outline of what we said. If spatially disjoint entities are to constitute a single spatiotemporal object, their interactions must be characterized by a certain regularity and predictability. Since it cannot be known without doing empirical research whether the various rocks in the world have this property, it cannot be said whether there is some "scattered object" consisting of all the rocks in existence.

But even if scattered objects do exist, properties aren't identical with them. Let R be some scattered object composed of all the rocks in the universe. If you destroy half of the rocks in the world, you have changed the just-mentioned scattered object. But you haven't changed the conditions that a thing must satisfy to be a rock. (To be a rock, a thing must still have a certain microstructure, or whatnot.) Therefore, you haven't changed what it is to be a rock or, consequently, the property of being a rock. So, by Leibniz's Law, that scattered object, if it even exists, cannot be identified with the property of being a rock. Given any other property, the same points *mutatis mutandis* show it not to be a scattered object.

2.3 Argument #3: The third-man argument

According to Platonism, any given spatiotemporal entity is an instance of at least one property, and for two spatiotemporal entities to resemble each other in some respect is for there to be some property of which each is an instance. Presumably, properties must resemble their instances. For example, the property of being triangular must itself be triangular. Bearing this in mind, let T be the property of being triangular, and let x be an arbitrary triangle. If Platonism is correct, it follows that there is some property T_1 such that T and x are both instances of T_1 and such that it is in virtue of their both being instances of T_1 that they are triangular. Given

> that, according to Platonism, properties resemble their instances, it follows that T_1 must resemble each of T and x. In other words, T_1 must be a triangle. It follows that, if Platonism is correct, there must be some property T_2 such that each of T_1 and T and x is an instance of T_2 and such that it is in virtue of their being instances of T_2 that they are triangular. And so on ad infinitum. Were they to exist, properties would be useless, since the very questions that arise in connection with their instances arise in connection with them.[55]

This argument evaluated: Consider the property of being spatiotemporal. If Platonism is right, that property is not itself spatiotemporal and, in the relevant sense of "resemble," it doesn't resemble its instances. Here is another example. To be triangular is to have a certain shape. Only things that are in space have shapes. The property of triangularity is not in space and, therefore, doesn't have a shape and, therefore, isn't an instance of itself. TMA succeeds in showing that instantiation isn't resemblance. This means that it succeeds in proving nothing, given that it's inherent in Platonism that instantiation *not* be resemblance.

2.4 Argument #4: the argument from the causal impotence of properties

Properties lack causal powers. For this reason, they can serve no explanatory function. To explain something is to show how it came about. Being causally impotent, properties can't make anything come about. Therefore, there is no reason to posit them.

This argument evaluated: Properties do indeed lack causal powers. Causes are necessarily in space and time. To be a cause is to induce change. Changes and their causes are *ipso facto* in space-time.

But this is no reason to think that properties are explanatory deadwood or, therefore, that there's no good reason to believe them to exist. Not all explanation is *causal* explanation. Some explanation is what we call *analytic* or *conceptual.* There are as many even numbers as there are whole numbers. (This is a theorem of set-theory.) How is this to be explained? Even though

the even numbers are a subset of the whole numbers, it is possible to put the two sets of numbers into a one-one[56] correspondence. Since, whenever two sets can be put into such a correspondence they have the same number of members, it follows that the set of even numbers has as many members as the set of whole numbers.[57]

This explanation is not *causal.* It explains a truth *not* by identifying its causal antecedents, but by delineating the structures of concepts. Many explanations are like this. You ask me: "Why can't someone know something without having any form of awareness?" I answer (correctly) by saying: "because knowing something involves having a true belief, and a belief is a form of awareness." This explanation identifies conditions that are conceptually, not causally, necessary conditions for knowledge. Like this explanation, Plato's theory isn't a causal theory, and it's therefore irrelevant that properties aren't in space-time.

2.5 Argument #5 (a variant of #4): the argument from the causal theory of knowledge (CTK)

> For x to know of y, y must have causal *effects* on x. You can see the dog only because the dog has certain effects on you—light bounces off its fur and hits your eyes. You can know about Socrates only because, long ago, Socrates made various noises, which were transcribed, and copies of those transcripts made their way to. So it is only because there is a causal process linking Socrates to you that you know about him and are able to have thoughts about him.
>
> Properties don't have effects on anything. They don't initiate or otherwise participate in causal processes. So we can't know about them, even if they exist. So there is no good reason to suppose that they exist.[58]

This argument evaluated: Seeing a dog does indeed involve the dog having effects on you—it involves your eyes being disturbed by photons bouncing off of the dog. But, in and of themselves, those photonic disturbances are lifeless and informationally empty. If they were to hit a rock, vision wouldn't

result. The photons responsible for those disturbances don't contain little dog images. You have to *read such information into them.* Your mind must extract such images from them; and to do this, your mind must have some way of interpreting those disturbances. The knowledge needed to interpret these disturbances cannot come from sense-perception, since that knowledge is itself a prerequisite for sense-perception. This entails that, in at least some cases, it isn't through a causal process that the knower becomes aware of the known. It also entails that a prerequisite for the very existence of sense-perceptions, or indeed of any mental activity that results from causal exchanges with the world, is knowledge of properties. Why does it entail this? Because, were it not for a grasp of what it is for a thing to have the property of having this or that shape, or being this or that distance from one, or having this or that color, one couldn't interpret the radiation-induced disturbances of one's sensory receptors in the way needed to generate sense-perceptions.

2.6. Argument #6: Conceptualism

> If they exist, properties are identical with mental representations of some kind or other—presumably, with thoughts *about* rocks or, in any case, with categories created by the mind that help it upload and process sensory information.[59]

This argument evaluated: Since conceptualism is thoroughly evaluated in Chapter 13, I'll be brief. There were rocks before there people. And, back then, rocks were similar to one another in ways that they weren't similar to bats and stars and other non-rocks For two things to be similar to one another in a certain respect is for there to be some property that they both have. So for two rocks to resemble each other in some respect is for there to exist some property that they both have. Thus, before there were human minds, there were properties. Therefore, properties are not products of human artifice.

2.7 Argument #7: Resemblance nominalism (RN)

> Let x be some red thing. x's being red consists, not in its being related in some way or other to some non-spatiotemporal entity but in its resembling roses, spots of blood, stop-signs, and other such things.

> Being red is a relation, not between the spatiotemporal and the non-spatiotemporal but between the spatiotemporal and the spatiotemporal. In general, to have a property phi isn't to bear some relation to some other spatiotemporal entity.[60]

This argument evaluated: For two otherwise dissimilar square-shaped objects to resemble each other in respect of shape is for there to exist some shape, and therefore some property, that they both have. For two otherwise dissimilar red objects to resemble each other in respect to shape is for there to exist some color, and therefore some property, that they both have. There is no such thing as *mere* resemblance; there is only resemblance in this or that respect. For two things to resemble each other in a given respect is for there to exist a property that they share, and resemblance nominalism therefore collapses.

2.8 Argument #8: Conventionalism

> The only non-empirical truths are linguistic conventions (e.g., "there are three feet in a yard"). Therefore, all non-conventional truths are empirical. Empirical truths register spatiotemporal facts. Conventions, whether linguistic or not, are obviously spatiotemporal entities. Therefore, all truths are spatiotemporal.[61]

This argument evaluated: Since conventionalism is evaluated in Chapters 1, 13, and 18, we can be brief.[62] Not all non-empirical truths are conventions. It's a non-empirical truth that there are continuous functions that cannot be differentiated at any point. But it's not a convention. *Given* the symbolic conventions that govern our usage of the relevant words ("continuous", "function," "differentiated", etc.), it *follows* that the sentence "there are continuous functions that cannot be differentiated at any point" expresses a truth. But this means that it holds in virtue of convention *plus* some various non-conventional facts. Those facts are not empirical; they don't concern the spatiotemporal world. Therefore, the fact described by that sentence is not spatiotemporal. Further, since that fact concerns logical structures of properties (e.g., the property of being differentiable), it follows that there are properties. Since, as we just noted, that fact is non-spatiotemporal, it follows

that properties are non-spatiotemporal.[63]

3.0 Nominalism

Many medieval philosophers, and some 20th century philosophers, held that properties are identical with words and, therefore, that properties are not spatiotemporal.

But this position is incoherent, since words and linguistic expressions generally are *themselves* universals or properties. How many words are right below this sentence?

“snow”

“snow”

“snow”

Is it one or three? The answer: there are three *instances* of one *word.* To put it another way, there are three *s tokens* of one word-*type.*[64] *Instances* of an expression are called expression-tokens and the expressions themselves are called expression-types.

No one utterance of the word “snow” is identical with that word. It’s not as though the word-type “snow” will disappear when the inscriptions of it in the line above this one is destroyed. That word-type is a *property* of certain deposits of ink (e.g., the ones above) and burst of noise, and those ink-deposits and bursts of noise are therefore instances of that property. This shows that, even if the nominalist is right to replace identify the property of being a rock with the word “rock,” he has still conceded the truth of Platonism.

Plus, that property isn’t identical with that word. The Spanish word for rock is “piedra.” It would be arbitrary to identify that property with the word “rock” but not with the word “piedra.” Since “rock” and “piedra” are obviously different words, the nominalist must say that the property of being a rock is not identical with itself, which is absurd.

All of these points aside, it makes no sense to identify that property, or any other, with a word. A thing’s being a rock has nothing to do with words; there were rocks before there were words.

4.0 Properties demystified

David Armstrong (1989) says that properties are “ways that things can be.” To be round is to be one way; to be square is another. Armstrong’s conception of what properties are is a fruitful one; and if it is kept in mind that properties are just ways that things can be, the resistances that we have to accepting their existence crumble.

Nobody denies that there are ways things can be. To say that there are ways things can be is to say that there are possibilities as to how things can

be. Not all of those possibilities are actualized. Not everything that *could be* the case is the case. Now, actualized possibilities are in space-time.[65] But *unactualized* possibilities— possibilities that never "came true"—are surely *not* in space-time. So where are they?

One could say that they don't exist. But that would be absurd. It's one thing to say that a possibility wasn't (or won't be) *actualized;* it's a very different thing to say that the possibility *itself* doesn't exist. It's possible that you will you go to law school. But (let us suppose) you're not going to go to law school (because you're going to pharmacy school instead). Does the fact that you *won't* go to law school mean that it isn't an option? Surely not. Your situation is very different from that of somebody (e.g., somebody on death-row) for whom it simply isn't an option. So the fact that a path isn't taken doesn't mean that the path doesn't *exist;* and the fact that a possibility isn't actualized doesn't mean that the possibility *itself* doesn't exist. So, given that possibilities, even unactualized ones, *do* exist, we must again ask: where are they? They're not in space-time. The *possibility* of your going to law school is not in space-time. So that possibility, being outside of space-time and thus being without location, is nowhere.

At first this may seem alarming, but it ceases to be so if we choose to see possibilities as properties and actualizations of possibilities as property-instances. It thus seems that the presumptive existence of unactualized possibilities provides support for the theory of universals.

A story may help clarify the nature of properties. You are a meat inspector. You've just inspected all the meat in slaughterhouse X. None of it passes inspection. All of it was completely rotten. There is obviously a way something *would* have to be to be a piece of meat that passed inspection; and there is obviously a way something would have to be to be a piece of meat *in slaughterhouse X* that passed inspection. So when reporting to your supervisor, you say "none of the pieces of meat in X deserved to pass inspection," you're saying: "none of those pieces of meat was the right way." Which is the same as saying: "there exists a way that something would have to be in order to be a piece of meat in X that deserved to pass inspection; but nothing is that way." Which is the same as saying: "the property of being a piece of meat in X that deserved to pass inspection is uninstantiated."

Thus, the property of being a piece of meat in X that deserves to pass inspection isn't a giant piece of meat in the sky. It's simply a way something would have to be in order to be a piece of meat in X that passed inspection.

The same is true of the property of being a pink elephant. There are no instances of that property, since there are no pink elephants. But there is a way that something *would* have to be to be a pink elephant. I know what that way is, and so do you. That's why we both know that, by going to the zoo and spray-painting one of the elephants there, we could *make* something be that way; and that's why we also know that nothing is that way (or, in any case, that nothing we've ever heard of is that way—maybe some prankster spray painted an elephant). For something to "instantiate the property of being a pink elephant" is simply for it to be that way. So that property isn't some pink elephant in the sky. It's just a way something would have to be to be a pink elephant. And for there to be no instances of that property is simply for nothing to be that way.

Properties can be thought of as conditions. The property of being a pink elephant is a condition a thing must fulfill in *order* to be a pink elephant. Lots of unfulfilled conditions exist. Mary won't marry someone unless he fulfills the following condition: he's a genius who is always friendly and never sleeps and always dotes on her and has a billion dollars. That condition isn't fulfilled by anyone. But the condition exists. Thus, properties that aren't instantiated are just unfulfilled conditions.

So given that there are unfulfilled conditions, it follows that there are uninstantiated properties. And given that properties exist, even when they have no instances, it follows that, when they do exist, they aren't identical with their instances. The property of being a pink elephant exists, but is currently uninstantiated. If there were to come into being a single pink elephant, which promptly died, the property wouldn't go out of existence. It would still be around. But if it were identical with instances, then it would go out of existence the moment its one instance did.

5.0 Relations and higher-order properties

We pointed out earlier that properties have properties. The property of being a bird has the property that there are more instances of it than the property of

being a penguin. Actually, any statement can be interpreted as a statement about higher-order properties. “Smith has a dog” is equivalent to “the property of having a dog has the property of instantiated by Smith.”

Higher-order properties are indispensable to scientific and ordinary discourse. If you say “true integrity is rare,” you’re saying that the property of having true integrity has the property of being seldom instantiated. If you say that ambition is more important to success than ability, you’re saying that the property of being ambitious has the property of comparing favorably with the property of having ability in respect to whether, in virtue of having it, its instances are likely to succeed.

Gottlob Frege (1844–1925) discovered that statements about numbers are statements about higher-order properties. “There are zero penguins in the house” means: the property of being a penguin in the house has the property of being uninstantiated. “There are two penguins in the house” says that the property of being a pair of penguins in the house is instantiated. The number two may be defined as what all pairs of objects have in common. For x to be a pair of objects is for it to be a set S such that there are objects O_1 and O_2, distinct from each other, that belong to S and such that anything O_2 that belongs to it is identical to either O_1 or O_2. The number two is the property of being such a pair. See Chapter 7.

Relations may be identified with properties of ordered pairs. There is some property P such that x is taller than y exactly if the ordered pair <x, y> has P. This conception of what relations are may seem contrived, but it’s an extremely useful one, since, without it, the logical forms of many inferences couldn’t be identified.

6.0 Sets and propositions identical with properties

We’ve seen that properties are not identical with sets of their instances. The main reason for this, we saw, is properties tend to be modally elastic, whereas sets are modally frozen. In other words, a given property (e.g., the property of being a poet) can generate different instances in different worlds, whereas the *set* of poets is frozen. If Smith is a member of the set of poets in this world, and S is the set of poets in this world, then S doesn’t exist in any world where

Smith does not (or where Smith does exist but isn't a poet).

But sets still may be identified with properties. Suppose that Smith, Jones, and Brown are the only three poets in existence. Let S be the set of actual poets—that is, the set containing Smith, Jones, Brown, and nothing else. In that case, the statement:

(1) x is a member of S

is equivalent with

(2) x is identical with Smith or Jones or Brown

and, therefore, with

(3) x has the property of being identical with Smith or Jones or Brown.

Thus, S may be identified with the property of being identical with Smith or Jones or Brown; and membership in S may be identified with possession of that property.

Sets are non-spatiotemporal entities. Given that sets are really properties, it follows that, given only that sets are non-spatiotemporal, it doesn't follow that anything non-spatiotemporal isn't a property.

Propositions are non-spatiotemporal entities. But they two are properties. The proposition *JM is playing tennis at time t* is true just in case certain properties are instantiated at certain times, and we may therefore identify that proposition with the set containing the relevant properties and identify that proposition's being true with the fact of all of that set's members' being instantiated. We'll work out the details in the next chapter. Since sets can, in the way just described, be identified with properties, it follows that propositions are properties.

Properties aren't mind-independent and neither (therefore) are propositions. But, of course, it's a popular belief, mainly among empiricists, that propositions, if existent, are human creations. Here is a reason, distinct from the one just given, to think otherwise.

There are infinitely many propositions. (Since there are infinitely many locations, there are, for any given particle x, infinitely propositions of the form *x is in place y at time t.*)We couldn't therefore have directly created every proposition there is. So, supposing that propositions are human

creations, we must be only *indirectly* responsible for infinity many of them, and it must therefore be supposed that we created some finite stock of propositions that somehow brought about the rest. But in that case there are true propositions, which we'll refer to as "intermediary" propositions, that describe how the ones we directly created brought about the others. But since there are infinitely many propositions that we didn't create, there are infinitely many intermediary propositions: one for each proposition not directly created. (Suppose for argument's sake that we didn't directly create the proposition that dinosaurs wore tweed, and also suppose that $p_1...p_n$ are the propositions that we directly created. In that case, there is a true intermediary proposition to the effect that $p_1...p_n$ brought about the propositions that dinosaurs wear tweed, another intermediary proposition another to the effect that $p_1...p_n$ brought about the proposition that *either* dinosaurs wore tweed *or* 1 + 1 = 2, and another to the effect that $p_1...p_n$ brought about the proposition that it's possible that it's the case that *either* dinosaurs wore tweed *or* 1 + 1 = 2, and so on.) We didn't create each of these individually many intermediary propositions. (Supposing that we didn't directly create the propositions that dinosaurs wore tweed; we also didn't directly create any proposition having it as a proper part—e.g., the proposition that *either* dinosaurs wore tweed *or* 1 + 1 = 2.) We must therefore posit a second group of intermediary propositions that will mediate between the first group and the propositions that we directly created. But everything just said about the first group of intermediary propositions will hold of the second, and we'll be no further along. Thus, it cannot coherently be supposed that we directly create every proposition in existence. This might appear to leave open the possibility that some propositions, but not others, are human artifacts. But such a view would be arbitrary and implausible, like the view that some whole numbers are human creations while others are not. It would also be demonstrably false. We've

seen that we don't directly or indirectly create the propositions describing how those that we do directly create, supposing that there are any, give rise to those we don't. Those that we don't directly create demonstrably entail those that we do. Suppose that we didn't create the proposition that dinosaurs wore tweed but that we did create the proposition that snow is white. In that case,

we didn't create the proposition that *dinosaurs wore tweed and snow is white.* That proposition entails that snow is white, and thus requires its existence. Therefore the proposition that snow is white doesn't need us to create it, since its existence follows from one that we didn't create. Therefore, we *couldn't* have created the proposition that snow is white, since one can't create what already exists.

Chapter 3

What Are Propositions and What Is Truth?[66]

1.0 Propositions as sets of properties, truth as instantiatedness of said properties

By a "proposition," I mean what is meant by an utterance of a sentence. If I point at Smith and say "that man is a scoundrel," the meaning of my utterance is is a proposition.

In this chapter, we'll provide answers to the questions "what are propositions?" and "what is it for propositions to be true?" Here's what we'll see. Propositions are ***sets*** of properties; and a proposition is true just in case the properties that are members of it are instantiated—so what it is for a proposition to be true is for the members of the set with which it is identical to be jointly instantiated. Roughly, truth is instantiatedness. Less roughly, truth is identical with the property of being a set (of a certain kind) all of whose members are instantiated. Aside from a few brief remarks, I will leave it open whether *all* sets of properties qualify as propositions (hence the parenthetical hedge in the last sentence). But a case will be made that *only* sets of properties are propositions.

2.0 The Argument

Consider the proposition:

1. Tommy is smoking.

Intuitively, this proposition seems to consist of at least three components: first, Tommy (or, in any case, something "individuative" of Tommy: perhaps a property that he uniquely instantiates or a concept under which he and he alone falls); second, the property of smoking; and, third, some kind of "synthesis" of the two—some complex constituent that involves Tommy's smoking. (We will try to state precisely what this synthesis involves.)

If 1 is true, then:

A. the property of being identical with Tommy

is instantiated, and so is

B. the property of smoking.

But, of course, those two properties could be instantiated in a world where Tommy is not smoking—in a world where, for example, Larry is smoking and a non-smoking Tommy is playing golf. We thus need a third property—a property such that, if a set S contains that property along with A and B, then 1 will be true just in case all of S's members are instantiated. That third property is not hard to identify. It is:

C. The property of being identical with something that is smoking and is identical with Tommy.

Before moving on, let's look at the apparatus that we've been developing. Once again, let S be a set that contains all and only A, B, and C[67]; and remember that we have identified S with 1 (the proposition that Tommy is smoking).

First of all, if C is instantiated, so are B and A. Further, if C, and therefore B and A, are instantiated, then, and only then, is 1 true. It follows that 1 is true if and only if each of A, B, and C is instantiated. This is consistent with, and even corroborates, the thesis that 1 *is* S and that 1's being true *is* S's members being jointly instantiated.

Of course, that does not by itself show that 1 can be identified with S or that 1's being true can be identified with S's members being jointly instantiated. After all, there are infinitely many distinct sets of properties such

that 1 is true just in case their members are instantiated. For example, let Z be the set containing the following properties and nothing besides: the property of being identical with a smoking Tommy; the property of being an even number; the property of being three sided *if* a triangle. The second two properties are instantiated in every possible world, given that even numbers necessarily exist and given that triangles cannot fail to have three sides. But surely Z cannot be identified with 1, the reason being that Z's membership cannot, at least not in a sufficiently natural manner, be aligned with what we know about 1's constituency.

But S's membership *does* so align, as a brief look back makes clear. S's members are (or at least include) A, B, and C. 1's constituents are (or at least include):

(i) something individuative of Tommy,

(ii) the property of smoking, and

(iii) something that combines those two things.

Obviously A is individuative of Tommy. (Only Tommy can have that property; and, what is more, he *must* have it.) So far, so good. (ii) and B are identical. So far, still so good.

What about C? That property is instantiated exactly if A and B are both instantiated. So, in a very clear sense of the term "combine," that property combines the property of smoking with (something individuative of) Tommy. (In the relevant sense of the term "combine," one property "combines" n other properties exactly if the first is instantiated exactly if the other n properties are jointly instantiated.) And we know that, in some way or other, 1 combines (something individuative of) Tommy with the property of smoking. In at least one respect, then, C's relationship to A and B is suggestively similar to 1's relationship to its parts. Moreover, C "dominates" each of A and B in a way suggestively similar to the way in which 1 dominates its constituents. A's being instantiated is necessary but not sufficient for C's being instantiated, the same being true of B's being instantiated. Tommy's existing is necessary but not sufficient for 1's being

true, the same being true of somebody's being a smoker.

While none of this demonstrates definitively that 1 is identical with S or that 1's being true is identical with S's being instantiated, we have, I believe, found enough of a match between 1/1's being true and S/S's members' being instantiated to warrant looking into the idea that they might be one and the same.

2.1 The unity of the proposition

How does 1 combine Tommy with the property of smoking? According to Frege, Tommy (or some concept thereof) *saturates* that property (or some concept thereof). (Henceforth, I'll omit the qualification "or some concept thereof.") But this doesn't help at all. For, as Davidson (1967) observed, the term "saturate" is just another label for the operation in question. So while the Fregean statement:

(FR) "1 is the result of Tommy's saturating the concept of smoking"

may well be correct, we don't know what it means.

We are in a position to give the meaning of FR and to say what this mysterious "saturating" amounts to. We've argued that 1 is a set whose members include the property of being Tommy, the property of smoking, and the property of being a smoking Tommy. 1 is the result of Tommy's "saturating" the property/concept of smoking in the sense that 1 is identical with a set S that consists of those three properties.

For any proposition X, to say that X is the result of Y's "saturating" Z (for some Y and Z) is to say that X is a set whose members are the property of being identical with X, and Z itself, and also the property of being identical with something that is identical with X and also has Z.

2.2 More on the decomposition of propositions

It is indisputable that 1 is true just in case C is instantiated. So why not just identify 1 with C, and 1's being true with C's being instantiated? Why go through the rigmarole that we went through? Why bother with A and B and, in general, any other properties other than C? Because if we identify 1 with

C, and 1's being true with C's being instantiated, our theory won't accommodate the very reasonable presumption that 1 has discrete parts corresponding to Tommy and to the property of smoking.

Consider an instance of the property of being a smoking Tommy. Obviously that instance will *involve* Tommy smoking. But it will not, at least in any obvious way, contain a discrete part corresponding to Tommy or a discrete part corresponding to smoking. And what is true of an instance of that property is, quite possibly, true of the property itself. It cannot be taken for granted that C has a *discrete* part corresponding to Tommy or to smoking or to anything else. In fact, being a non-spatiotemporal entity, C won't have discrete parts in any straightforward sense at all.

The impression that C is part-less, or at least lacking in the right kind of parts, is reinforced when we look at what an instance of C is like. Such an instance is, to adapt Kenneth Taylor's (1998) apt expression, a dent or ripple in the quantum; it is a shift in the movement of mass-energy. Of course, that mass-energy is (by hypothesis) sufficient for the existence of Tommy and of smoking and, indeed, of a smoking Tommy. But you would have a devil of a time isolating some part of that distribution that corresponded to Tommy, then distinguishing that from some other part corresponding to smoking, and then finding a third that fused those two, otherwise discrete, entities into a discrete instance of a smoking Tommy.

By contrast, the proposition:

1. Tommy is smoking

is, of its very nature, neatly parsed into Tommy, the property of smoking, and a third constituent that synthesizes the two just mentioned. The neat articulations found in 1 are found, in some form, in S. Those articulations, though found in S's membership as a whole, are not found in C taken by itself, making S, but not C, a suitable mirror for the decompositional structure of 1.

States of affairs (or "facts," as they are sometimes called) do not, as Ian Hacking (1975) long ago put it, have the same kinds of "articulations" as propositions. The quantum-ripples on which the truth of a proposition

supervenes don't have anything like the structure of that proposition. Whereas the proposition *Tommy is smoking* clearly has a unique, and relatively simple decomposition into Tommy (or some property/concept individuative thereof), the property of smoking, etc., the state of affairs in which Tommy's smoking consists—the confluence of psychological, physiological, and, ultimately, sub-atomic disturbances—obviously doesn't uniquely decompose into Tommy, the property of smoking. In fact, it seems a stretch to suppose that it even has *one* such decomposition.

2.2.1 A corollary

Consider the proposition:

2. Tommy is snoring.

Let D be the property of being a snoring Tommy. It is obvious that:

1. Tommy is smoking

and

2. Tommy is snoring

have much in common. They have in common the constituent *Tommy* (or, in any case, something corresponding thereto). They have in common the property of being a Tommy who is doing something (in other words, they have in common the property of being identical with a thing x such that x is identical with Tommy and such that, for some y, x is doing y); and they also have in common the property of being a thing that is doing something (in other words, they have in common the property of being a thing x such that, for some y, x is doing y).

In light of this, consider some instance of D. In other words, consider the kind of property-instance, the kind of mass-energy distribution, on which the truth of 2 supervenes. That instance will, of course, suffice for the existence of Tommy, as will an instance of C. But will those instances, in virtue of that fact, have some *discrete* part in common? Unlikely. In neither case do we have Tommy *simpliciter.* We have *Tommy smoking (while sweating profusely and worrying about his algebra test and his nascent nicotine addition, all the*

while talking with his cousin Albert); in the other we have *Tommy snoring (while having terrible dreams about his future as a second string pitcher on a mediocre team).* There is no such thing, in either case, as Tommy *simpliciter.* There is only Tommy doing this or that, having this or that property. A corollary is that two situations that comprise Tommy do not, in virtue of that fact, have in common some discrete, isolable entity. So while it is true that C and D both comprise Tommy, the one does not, in comprising him, have a discrete constituent in common with the other.

But now we must consider what we've said about the composition of 1 and also what follows, by obvious extensions of what we've said about 1, about 2. 1 is a set S whose members are:

a: the property of being Tommy

b: the property of smoking

c: the property of being a smoking Tommy

And, by similar reasoning, 2 is a set S* whose members are:

a: the property of being Tommy

b*: the property of snoring

c: the property of being a snoring Tommy

S and S* *do* have well-defined, discrete parts in common. Indeed, they have in common precisely what is had in common by:

1. Tommy is smoking

and

2. Tommy is snoring.

If we were to be really precise about it, we'd say that each of S and S* comprised, not only a–c, but also

d. the property of being a Tommy that does something (i.e., the property being a Tommy that does something or, in any case, has some property),

and also

e. the property of being a something that does something (i.e., the property being a something that does something or, in any case, has some property).

Let's look at 1 for a second. It immediately and formally implies:

i Tommy smokes;
ii something smokes;
iii Tommy does something;
iv Tommy exists;
v something does something.

And 2 immediately implies:

i* Tommy snores;
ii* something snores;
iii Tommy does something;
iv Tommy exists;
v something does something.

Now let's look at S and at S*. The members of S are:

I The property of being Tommy
II The property of smoking
III The property of being a Tommy that smokes
IV The property of being a Tommy that does something
V The property of being a something that does something

The members of S* are:

I The property of being Tommy
II* The property of snoring
III* The property of being a Tommy that snores
IV The property of being a Tommy that does something

V The property of being a something that does something

Having identified 1 with S and 2 with S*, everything we know about those propositions—about what they entail, about what their components are, and about how their higher-level components are "forged" out of their lower level components—is accounted for.

2.3 Another desideratum satisfied by our analysis

Many semanticists today hold that Plato and Socrates are *veritable* constituents of the propositions that concern them. So according to Kaplan (1989),who is following and agreeing with Russell (1903), Timmy—the person, not some Fregean concept thereof—is a constituent of both 1 and 2. Other adherents of this view are Scott Soames (2003), Nathan Salmon (2005), and Robert Moore (1995). (Opponents are J. Searle (1983), M. Dummett (1973), and—with some very heavy reservations to be stated later—the present author.)

There is a rather serious problem with this view. Consider, not 1 or 2, but rather:

3. Plato snores.

At this point in time, Plato doesn't exist. So if he's a (necessary) constituent of 3—if, in other words, 3 exists only if Plato does—then 3 doesn't exist. Salmon and Soames, both of them hard-line direct reference theorists, deal with this by taking a desperate measure, viz. by saying that 3 exists *and* that its existence constitutively depends on Plato's existence *and* that these two positions can be reconciled with each other—that, more specifically, they can be reconciled by saying that 3 exists, albeit as a non-existent proposition.

This is not necessary or plausible or helpful. 3 exists. Plato does not. But the property of being identical with Plato, whatever it is, does exist. What is it to be Plato? To be Plato is, perhaps, to be a causal sequence that is initiated by the fertilizing of a certain egg by a certain spermatozoa. To be Plato is, perhaps, to have a soul or mind of a certain kind. What we know is that, if something is Plato, its being Plato supervenes on something probably having to do with the mass-energy distribution in the cosmos—on its mode of

origination, on the quotient of spiritual energy in the particle-interactions that constitute it, on who knows what. In any case, even though Plato does not exist, the property of being identical with Plato does. That's why I can meaningfully, not to mention correctly, say "I am not Plato" or, equivalently, "I don't have the property of being identical with Plato." And it's also why we can meaningfully, and presumably correctly, say that Plato does not exist. For, in making such a statement, we aren't picking out a non-object—which would, having been picked out, *ipso facto* cease to be a non-object—and then saying of it that it doesn't exist. We are saying of an existing property, viz. that of being identical with Plato, that it is uninstantiated.

But our analysis is still consistent with the heart of direct-reference-theory, this being the idea that, for some x such that x is identical with Plato, the proposition meant by "Plato snores" is true just in case x snores, it being completely irrelevant what properties (other than being a snorer) x has. (The Fregean indirectreferentialist holds that for "Plato snores" to be correct is necessary that, for example, a unique great philosopher to die of hemlock-poisoning snore.) It is clear why our analysis is consistent on this matter with the viewpoint of the direct referentialist. According to our analysis, 3 is true exactly if the following properties are instantiated:

a. the property of being identical with Plato

b. the property of snoring

c. the property of being a Plato who snores

d. the property of being something that snores

e. the property of being something that does something (or has some property)

It is readily seen that, on our analysis, there is some x, such that x is Plato and such that 3 is true exactly if x snores, it being irrelevant what other properties x has. This is exactly what the direct-referentialist holds. At the same time, our analysis guarantees the existence of the proposition *that Plato snores*—after all, the *property* of being Plato exists, even though it is now uninstantiated. So our analysis of *that Plato snores* gives the direct-reference-theorist the truth conditions that she wants and that (given certain compelling

arguments due to Kripke 1980 and Soames 2003, 2004) she probably deserves. At the same time, our analysis doesn't put the direct-reference-theorist in the awkward, logic-bending position of having to say that the proposition that Plato snores exists despite the failure of existence of entities on whose existence its own existence is constitutively dependent.

3.0 Molecular propositions

(Because this section and the three that follow it are more technical than substantive, I suggest skipping straight to Section 4.0 on a first and probably a second reading.) Our analysis applies to molecular propositions no less than to atomic propositions. But making this transposition go through will involve a certain amount of sensitivity to details and technicalities. A certain amount of "engineering," as opposed to philosophy "proper," will be needed.

3.1 Existence-claims

In keeping with tradition, I will treat generalizations as molecular propositions. So:

4. something snores

is an existential generalization. Given what we've said, the most obvious analysis of 4 would be to say that 4 is the property of snoring (or a set whose sole member is that property) and that 4's being true is identical with that property's being instantiated.

That would not be the correct analysis. Even though 4 is true just in case the property of snoring is instantiated, 4 is not identical with that property and its truth is not identical with that property's being instantiated. We know from Frege that 4, unlike the property of snoring, has a rather distinctive decompositional structure, and that 4 thus consists of (at least) three or four constituents. (Note: In this context, by an "instantiated function," I mean a proposition-form at least one of whose members is correct. Thus x snores is such a function. An "instance" of a function is anything satisfying it. So, supposing that he snores, Smith is an instance of the just-mentioned function. A "propositional form" may be thought of as a class of propositions all of the

same form.)We also know from Frege what the first two constituents are:

a. the property of snoring

b. the property of being instantiated

In light of what we've already said, it is not hard to generate some (if not all) of the remaining constituents of 4, these being:

c. the property of being an instantiated function each of whose instances snores

d. the property of an instantiated function each of whose instances has some property or other

Let S be a set containing all and only a–d. 4 will be true iff a–d are instantiated. So, from the viewpoint of truth-conditions, there is no barrier to identifying 4 with S or to identifying 4's being true with S's members being instantiated. Further, 4's structure is easily mapped onto S's. 4, if true, entails that there exists some property that has some instances; that there exists at least one snorer; that at least one property is instantiated; and that at least one property exists. And all of this is precisely what S's members' being instantiated involves.

Now consider the proposition:

5. Something smokes.

By reasoning similar to that just given, the constituents of this proposition are:

a.* the property of smoking

b. the property of being instantiated

c.* the property of a function each of whose instances smokes

d. the property of a function each of whose instances has some property or other

We know pre-theoretically that 4 and 5 have certain constituents in common (e.g., the property of being instantiated or, in any case, *some*

property expressed by "something"). Our analysis of 4 and 5 validate everything we know, theoretically and pre-theoretically, about the decompositions and logical forms of the propositions.

3.2 Negative non-quantified propositions

Before considering "nothing snores," I want to consider a negative, but non-quantified, proposition:

6. Timmy is not smoking,

in other words,

6. It is not the case *that Timmy is smoking.*

(The reason for the underlining will be clear in a moment.) The two immediate constituents of 6 would *appear* to be "It is not the case" and "that Timmy is smoking." But there's a problem here. We have said that propositions are sets of properties that are true when instantiated. In any case, this is what we have said about atomic propositions and, for uniformity's sake, this is what we must say about molecular propositions. But if we say that, in 6, the underlined proposition is instantiated—that is, that all its members are instantiated— then *that Timmy is smoking* will be true and 6 will be false and indeed self-negating.

Not-P is true just in case P is false. So "not" can be thought of as expressing a function that assigns truth to all and only false propositions or (what, in this context, is scarcely different) as a property of all and only false propositions.

For the reason given a moment ago, we don't want to say that *that Timmy is smoking* is a constituent of 6, and will instead say that the property of being *identical with* that proposition is such a property. We will also say that the property of being a false proposition is such a constituent. Then we synthesize those two components into the property of being the false proposition that Tommy is smoking. That property, it will be noticed, is instantiated exactly if the proposition that Tommy is smoking is false. So, according to our analysis, the constituents of 6 are:

a. the property of being a false proposition

b. the property of being identical with the proposition *that Tommy is smoking*

c. the property of being identical with a proposition that is false and that is identical with *that Tommy is smoking.*

Let S be the set containing all and only a–c. Let us consider what happens when each of S's members is instantiated. First, suppose that c is instantiated. In that case, there is an instance of a proposition that is false and that is identical with the proposition that Tommy is smoking. So 6 is true just in case c is instantiated. Since, if c is instantiated if the other members of S are instantiated, it follows that 6 is instantiated if and only if each of the members of S is instantiated. Thus, there are, as of yet, no bars to identifying 6 with S and 6's truth with S's members' being instantiated.

Next, suppose that b is instantiated. In that case, the property *Tommy is smoking* exists—but it isn't necessarily true. And if 6 should be true—which, according to our analysis, would mean that all of a–c were instantiated—we'd be stuck with the *existence,* but not the *truth* of the proposition that Tommy is smoking. Which is precisely what we want.

Finally, suppose that a is instantiated. In that case, the property of being a false proposition is instantiated. And, of course, if 6 is true—if, in other words, the members of S are all instantiated—then that property is instantiated. Which, again, is precisely what we want.

Thus, our analysis is consistent with most of our pretheoretic intuitions about the decomposition of 6. (The one exception that being that, where our pre-theoretic intuition wants to *that Tommy is smoking* in 6, our analysis puts the property of being identical with that proposition in its place.) And our analysis is also inconsistent with the logical properties of 6.

3.3 Quantified propositions revisited

Consider:

7. Everything snores.

By reasoning similar to that already given, the main constituents of this are:

a. the property of snoring
b. the property of being universally instantiated
c. the property of being identical with a universally instantiated property that is identical with the property of snoring

Of course, c is a property each of whose instances is an instance both of (a) the property of snoring and (b) of the property of being universally instantiated.

Let us now consider:

8. Nothing snores

I will treat this is as the negation of 5. I will treat it, in other words, as:

8: It is not the case *that something snores.*

For reasons now already clear, the constituents of 8 include (but are not confined to):

a. the property of being a false proposition
b. the property of being identical with the proposition that something snores
c. the property of being identical with a false proposition that is identical with the proposition that something snores

Notice that, if c is instantiated, so are b and a. And notice that, if c is instantiated, then the proposition that something snores is false. Finally, notice that c is related to a and b in the usual way—any instance of 6 is an instance both of (a) the property of being a false proposition and (b) the property of being identical with *that Tommy smokes.*

3.4 Non-quantified molecular propositions

Let's finish up by considering two non-quantified molecular propositions. Consider the sentence:

9. Tommy is smoking or Tommy is snoring

9 can be seen as saying of a certain pair of propositions (namely, *that Tommy*

is smoking and *that Tommy is snoring*) that at least one of its members is true. I thus see 9 as consisting of:

a. the property of disjunction

which in this context we may regard as the property of being a pair of propositions at least one of which is true, and

b. the property of being identical with a pair of propositions, such that one member of that pair is *that Tommy is smoking,* the other member being *that Tommy is snoring.*

Of course, 9 could be false even if both a and b were instantiated. (This would be the case in any world where both *Tommy is smoking* and *Tommy is snoring* were false, but where some disjunction or other was true.) So, for now familiar reasons, we must look for a third constituent, where that constituent must be a property c such that each instance of c is (i) an instance of a pair of propositions at least one of which is true and is also (ii) an instance of a pair of propositions, one them being *that Tommy is smoking,* the other being *that Tommy is snoring.* So our final properties are:

a. the property of being a pair of propositions, at least one of which is true

b. the property of being a pair of propositions, one of them being *that Tommy is smoking,* the other being *that Tommy is snoring*

c. the property of being a pair of propositions, at least one of which is true, and at least one of which is *that Tommy is smoking* and at least one of which is *that Tommy is snoring*

Of course, if c is instantiated, then so are b and c. And if c is instantiated, then either Tommy is snoring or Tommy is smoking. So c, and therefore a–c, are instantiated exactly if 9 is true. So if S is the set containing a–c, then 9 is true iff all of S's members are instantiated.

We've thus made it clear (i) what it is for something to be a "constituent" of a proposition, (ii) what it is for an argument to "saturate" a function, (iii) more generally what it is for two propositional constituents to "synthesize" into a third constituent, (iv) what it is to be a proposition, and (v) what it is for a proposition to be true. And our analysis is to a non-trivial degree

(though not completely) consistent with what pretheoretic intuition and common sense have to tell us about propositions.

4.0 The Russell-Kaplan-Salmon analysis of propositions

According to Russell (1903), Salmon (1986), Kaplan (1989), R. Moore (1995), J. Perry (1996), J. King (1996, 1997), to name but a few, a proposition is some kind of "structure" consisting of "objects and properties." So, according to this view, the proposition expressed by (tokens of) "Socrates is bald" is some kind of structure consisting of Socrates—the man himself, not some property that he has or some concept that he satisfies—and the property of baldness. It is usually said that the structure in question is an *ordered* pair. At least part of the reason for is that different propositions seem capable of having the same parts (cf. *Brutus killed Caesar* and *Caesar killed Brutus*). And unless the sets with which propositions are identified have some kind of ordinal structure, distinct propositions (e.g., the two mentioned in the last parenthesis) will collapse into one another. Thus *Socrates is bald* is identified, not with (Socrates, baldness), but instead with <Socrates, baldness> or with the <relation of having <Socrates, baldness>>, or some such.

4.1 Problems with this view

The ordered pair <2,7> isn't true or false. Neither is the ordered pair <2,7 <less than>>. Given that these ordered pairs aren't true or false, why is <Socrates, the relation of having <baldness>>rue or false? There's no denying that we could *use* the corresponding expression (viz. "<Socrates, the relation of having <baldness>>") to express the proposition expressed by the sentence "Socrates is bald." We could obviously invent a code whereby "<the relation of having <Socrates, baldness>>" meant *Socrates is bald.* But this fact tells us little, if anything, about the structure of the corresponding proposition that we didn't already know from the fact that it can be expressed by "Socrates is bald."

4.1.2 Another concern

Obviously the brackets are meant to be symbolic of *some* kind of order that holds among Socrates and baldness (and the relation of having). But what is that ordinal relation? If that ordinal relation bears no resemblance to the relation that would *actually* be had by Socrates, in a situation where he was bald, with respect to the property of baldness, then it wouldn't be clear why what relation was the right one: some kind of reason—the likes of which have never been provided by advocates of this model—would have to be provided why that particular ordinal relation was the right one. So, thus construed, the analysis in question is highly programmatic. On the other hand, if that relation *is* the very relation that would be had by Socrates, with respect to the property of baldness, in a situation where he was bald, then Socrates' baldness—his actual baldness—would be built into the very existence of the proposition that he was bald. In other words, the proposition that Socrates is bald would, by virtue of its own configuration, make him bald, and would thus predetermine its own truth. But obviously the truth of *Socrates is bald* isn't self-determining. Basically, if a theory requires propositions to model facts *too* well, then that theory wrongly requires contingent propositions to be true; but a theory runs the risk of arbitrariness if the relations that, according to it, hold among a proposition's parts are not in any intuitive way counterbalanced by the facts that, if that proposition were true, hold among the components of the corresponding fact. If we turn the proposition *Socrates is bald* into a model consisting of Socrates standing in the relation of having with respect to baldness, then the proposition *itself* is the fact of Socrates' being bald, and no way is left open for the obvious possibility that it might be false. At the same time, it would be easy to render overly tenuous and schematic the relationship between, on the one hand, the fact of Socrates' being bald and, on the other hand, the relationship that, in the proposition that Socrates is bald, Socrates bears with respect to baldness. This mistake is, I believe committed by the theory according to which that proposition is identical with ordered pair <the relation of having <Socrates, baldness>>. The relation borne that, in that structure, Socrates bears with respect to baldness is, it seems to me, too schematic—too without any counterpart in the fact that, supposing that proposition true, makes it so.

4.1.3 Another problem

Though it is now false, the proposition *Socrates is alive* obviously exists. What if, unbeknownst to you, your friend Joey died some time during the last five minutes? Surely you wouldn't want to say that the proposition *Joey loves to read* suddenly ceased to exist. Apart from an acceptance of a certain version of direct-reference theory, there is no reason to accept that view.

Some authors (e.g., Salmon (2005)), *do* say that *Joey loves to read* does cease to exist when Joey dies (or that it continues to exist, but only in some curious form—that it exists, but only as a non-existent entity, or some such). The reasoning underlying this extraordinary view seems to be as follows. There is some x—namely, Joey—such that "Joey loves to read" is true exactly if x loves to read. It follows, it is said, that the proposition meant by that sentence—namely, *Joey loves to read*—must have Joey as a constituent. And from this, it is said, it follows that *Joey loves to read* ceases to exist when Joey does.

This reasoning is spurious. If only for argument's sake, let us suppose that, for some x, such that x is Joey, "Joey loves to read" is true iff x loves to read. It doesn't follow that *Joey loves to read* ceases to exist when Joey does. For, if anything like the view I am advocating is right, it could be that it is not Joey *per se,* but the property of being identical with him, that is a constituent of that proposition. Our view is consistent with the idea that "Joey" rigidly designates Joey—that there is some x such that x is Joey such that "Joey loves to read" is true just in case x loves to read, it being irrelevant what properties are had by other things. But, according to our view, it isn't Joey *per se,* but some property that is a constituent of *Joey loves to read.* That property, which is the property of being identical with Joey, is one that nothing other than Joey could possibly instantiate. (It isn't like the concept *bifocal inventor,* which could be instantiated by something other than the thing which is its actual instance.) According to our analysis, the proposition that Joey loves to read is *true* only if that property is instantiated. Thus, there is indeed some x, namely Joey, such that each of *Joey loves to read* and "Joey loves to read" is true iff x loves to read. But Joey's existence is a pre-requisite for the truth, not the existence, of that proposition. So, unless one maintains (implausibly) that the property of being identical with Joey can exist only if it is instantiated, it follows that *Joey loves to read* doesn't cease to exist when Joey does.

Like the present author, Frege held that it wasn't Socrates *per se* that was a part of the proposition *Socrates is alive* but was instead some concept that Socrates uniquely satisfied. (Frege's analysis is therefore consistent with the fact that "Joey likes to read" holds onto its meaning even if Joey ceases to exist.) But, apart from that, Frege's view is quite as vague as Russell's. Frege was extremely opaque about the nature of the structure of this proposition; and what he failed to provide in the way of clear formulations he tried, unsuccessfully, to make up for with unhelpful terms such as "saturate."

What is good about the Russell-Frege-Kaplan-King view is that it insists that propositions resemble the sentences that express them. So that view is, at least, consistent with the fact that languages can be learned and used. (For, presumably, learning and using languages involves there being systematic ways of hooking up meanings to sentences.) And, as we'll now see, other analyses don't do such a good job of satisfying this requirement.

5.0 Propositions as sets of their own consequences

For a while, it was widely held (by e.g., Clarence Lewis (1946) and Rudolph Carnap (1934)) that the "intension" of a sentence (which, in this context, we may identify with what we would now call the "meaning" of that sentence or the "proposition" expressed by it) was identical with the class of all propositions that could be inferred from that sentence. So (a natural extension of this view would be that) the proposition *John knows that corn is yummy* is the class of all propositions deducible from that proposition—for example, *John believes that corn is yummy, John knows something, John has some kind of attitude towards the proposition that corn is yummy.*[68] There are a few problems with this analysis, some minor, some major. A minor problem is that, as it was stated by Lewis and Carnap, analysis is viciously circular, since many propositions can be deduced from *John knows that corn is yummy* that involve that very proposition—for example, *either John knows that corn is yummy or there are square circles.* This particular problem is not, I believe, insuperable, since the position can be amended to avoid it. The right emendation would seem to be: a proposition is identical with the class of all propositions—setting aside that proposition itself, and any propositions

having it as a component (e.g., disjunctions of which it is a disjunct)—that can be inferred from it.

5.1 Why this view is false

Even so, the position in question isn't very plausible. Consider the class of all propositions that follow from *John knows that corn is yummy.* That class doesn't have a structure corresponding even remotely to the structure that we would expect that proposition to have, given that it almost certainly has a structure non-trivially similar to that of the sentences that express it. Also, in order to figure out whether a given proposition is a consequence of *John knows that corn is yummy,* one must grasp that proposition itself. Here's why: The set of propositions that follow from *John knows that corn is yummy* is infinitely large. One can grasp that set only in the sense that one grasps some rule that generates its members—after all, one obviously cannot grasp them all one by one. And grasping that rule would seem to involve grasping *John knows that corn is yummy* or some other similar proposition.

6.0 The possible worlds approach

We now turn to what is nowadays probably the most popular analysis of propositions and, indeed, of modality: the possible worlds analysis. According to possible worlds semantics (PWS), a proposition is necessary (possible) if true in all (some) worlds, and a proposition is a function that assigns truths–values to worlds or, alternatively, is a set of worlds (viz. the set of worlds where it is true). So the proposition

grass is green is a function that assigns *truth* to all the world where grass is green, *falsity* to worlds where grass is not green, and either assigns falsity to grass-free worlds or assigns them no truth-value at all.

In this context, the first question to ask is: what is meant by the term "world"? Oftentimes, advocates of PWS say that "worlds," as they are using the term, are propositions (or sets of propositions). In that case, as David Lewis (1986) was quick to point out, the PWS slogan that "a proposition is a set of worlds" becomes "a proposition is a set of propositions," which is not only false but viciously regressive, the same thing *mutatis mutandis* holding of the statement "a proposition is a function that assigns truth-values to worlds."

Aware that treating worlds as propositions (or sets thereof) eviscerates PWS, some advocates of that doctrine say that worlds are just that—worlds: concrete objects like our own, "maximal states of affairs." *Not,* it must be emphasized, properties that, if instantiated, *would* be worlds. But veritable worlds.[69]

6.1 Problems with this approach

Not too many are going to grant the existence of *bona fide* alternative worlds. Second, even if they do exist, we can't know anything about them *except what we can figure out on the basis of modal knowledge obtained independently of our knowledge of those worlds.*

Lewis (1986) says, correctly, that we don't know about alternative worlds by taking a tour of them. We know about them by considering materials that are available to us here and "employing a principle of recombination"—by, basically, applying modal principles that we already know to hold and then seeing if those principles square with the proposition whose modal status we wish to know (e.g,. there might be square circles in other worlds). But in that case, the worlds answer to us, not us to them. Our modal knowledge is the tail that wags the dog, the dog being these other worlds; and from the viewpoint of explaining (or validating) our modal knowledge, those extra worlds are useless (though there may be other good reasons for positing them).

6.2 Other problems

The class of worlds where 1 + 1 = 2 is identical with the class of worlds *triangles have three sides.* But those propositions are obviously distinct, that being why you can think the one without thinking the other. (There's no intensional fallacy here, since propositions *are* intensions.) In its most primitive form, PWS can't distinguish analytically equivalent propositions. Some advocates of PWS deal with this by saying the following:

> There are two ways of thinking of what functions are. A function can be thought of as a set of ordered pairs. So, from that viewpoint, the function F(x) = x + 1 just is the set of pairs: <1, 2>, <2, 3>....And from that viewpoint, F(x) = x + 1 is identical with the function F(x) = x + (90 – 89). This is the *extensional* view of functions: functions just *are* the pairs they generate—*how* they generate them is irrelevant. But the extensionalist view of functions isn't the only one, since there's also the *intensionalist* view. The intensionalist says that functions are individuated by *how* they generate the pairs they generate. So, for the intensionalist, for functions to be identical, it is necessary but not sufficient that they generate the same ordered pairs. To avoid falsely identifying distinct, analytically equivalent propositions, PWS must take an intensionalist view of functions. So propositions are functions, and those functions are individuated by their intensions, not their extensions.

There are two problems here. First, this view still demands that there be other worlds. If those worlds are propositions, then we're no further along. If those worlds are concrete, then we're stuck with some useless metaphysical lumber. But leaving those problems aside, the solution just proposed collapses on itself. Propositions *are* intensions (though not all intensions are propositions—e.g., the propositional function *x is a human being* is not a proposition, even though it is an intension). This by itself is a problem for PWS since much of its *raison d'être* is that it replaces intensions, which for various reasons are seen as inimical to scientific thought[70], with extensions, which are seen as methodologically unimpeachable. So in taking an intensionalist view of

functions, PWS is undermining itself. For, in taking such a view, PWS is letting properties back in—properties are no longer to be identified with sets of their instances. But, by parity of reasoning, propositions needn't be identified with sets of *their* instances (i.e., with sets of worlds where they hold). So PWS cannot, without arbitrariness, take an intensionalist view of functions; but it cannot, without failing to distinguish between obviously distinct propositions, take an extensionalist view of them either.

6.2.1 A similar problem

Consider the proposition *grass is green.* As we know, PWS identifies that proposition with a function (that assigns truth to worlds where grass is green) or with a set of worlds (containing all and only those worlds where grass is green). If it isn't thought of as a set of ordered pairs, that function must be thought of as a *rule* of some kind. What kind of rule? One that assigns truth to certain worlds on the basis of their satisfying some condition. What condition is that? The condition that, in those worlds, grass is green. So the function in question, understood intensionally, must be understood *in terms of the very proposition in question.* So unless the advocate of PWS wants to falsely identify 1 + 1 = 2 with *triangles have three sides,* he must take an intensionalist view of functions, which means that he must circularly understand the functions with which he aspires to identify propositions with the very propositions with which he aspires to identify them.

7.0 Conclusion

Frege (1918) made it clear that propositions are the *contents* of psychological entities and are not themselves psychological entities. Since Frege, many attempts have been made to say what propositions are. Those attempts have either been programmatic to the point of near-emptiness or demonstrably false or both.

But it is possible to identify structures that model the properties that we know, on both theoretical and commonsensical grounds, propositions to have. Given any proposition P, it is clear that there is some set of properties $p_1...p_n$ that will be instantiated just in case P is true. Supposing that S is the set

containing those properties, this suggests that S *is* P and that P's being true *is* S's members' being instantiated.

We found that, taking this approach, we could give non-metaphorical meanings to terms that, though liberally used, have thus far been used in a purely metaphorical way—terms like "constituent" and "synthesize." A "constituent" of a proposition, we found reason to believe, is a *member* of the set (or, more accurately, of *some* set, since there are many of the kind in question) of properties whose joint-instantiatedness is necessary and sufficient for that proposition's truth. As for the term "synthesize": since all components of propositions are properties, only properties can, in any relevant sense, be synthesized. And two properties p_1 and p_2 are "synthesized" into another component when a third property, p_3, is identified each of whose instances is *ipso facto* an instance of both p1 and p_2. (Or, to put it in a way that doesn't have any misleading chronological or psychological overtones: p_3 is a synthesis of p_1 and p_2 just in case any instance of p_3 is *ipso facto* an instance of each of p_1 and p_2.)

PART II

Language and Thought

Chapter 4

What is Language?

1.0 The meaning of "meaning"

There would be no languages if there were no expressions (words, phrases, sentences, etc.). Nothing meaningless is an expression. For this reason, the concept of an expression must be understood in terms of the concept of meaning, the same therefore being true of the concept of language.

But it isn't much use to be told that words and sentences "have meanings," since the word "meaning" has three different meanings, and only one of these directly relates to the nature of language.

Meaning #1: The evidential meaning of "meaning"

In some cases, to say that x "means" y is to say that x is *evidence* of y—that x and y are causally interrelated in such a way that, given x, it can reasonably be inferred that y. "Smith's hacking cough means that he has a violent lung infection" means "Smith's hacking cough is evidence that he has a violent lung infection." And the latter means that coughs like Smith's are *causally connected* to violent lung infections in such a way that it may reasonably be inferred that Smith has a violent lung infection.

Smith's violent lung infection is a *cause* of Smith's hacking cough. But for x to be evidence to of y, it is neither necessary nor sufficient that y cause x. Why isn't it *necessary?* First of all, causes can be evidence of their own effects. (Bill's current drunkenness is evidence of a poor performance on his upcoming economics exam.) Second, if some event or state of affairs z is a cause of both x and y, then x can be evidence of y without being a cause or an effect of y. (Suppose that Bill is slurring his words. This is evidence that he'll do poorly on the upcoming test. But his slurring his words is neither a cause, nor an effect, of his substandard test-performance. His drunkenness is a cause of (a) his slurred speech and (b) his imminent, substandard test-performance. And it's because his slurred speech has a causal ancestor in common with his

poor test-performance that the former is evidence of the latter.)Why isn't it *sufficient?* Given only that the cause of Bill's failing was that he was drunk, we can't infer from the fact that he failed that he was drunk. There are many reasons why a person may fail a test. Supposing that y caused x to occur, x is evidence of y only if it can't reasonably be supposed that anything other than y was the cause. (Only a violent lung infection could be responsible for Smith's hacking cough, that being why the latter is evidence of the former.)

Meaning #2: The psychological meaning of "meaning"

When we use sentences, we mean things by them. Meaning in this sense is a psychological notion. You tell me that Sally is the most wonderful, decent person you've ever known. I respond by saying "things aren't always as they seem." What I mean is that Sally is devious. In other words, my intention in making this statement is to say that Sally is devious. Given that intentions are psychological entities, meaning in this sense is obviously a psychological notion.

Meaning #3: The linguistic meaning of "meaning"

The sentence "snow is white" says something about something; it attributes the property of being white to snow. Therefore, it *means* that snow is white. This kind of meaning is in a class by itself; it isn't identical with either of the two kinds mentioned so far; and there isn't any obvious way to understand it in terms of them. Let us now say why.

Meaning in the psychological sense *involves,* but does not coincide with, meaning in the linguistic sense. Once again suppose that, in response to your telling me that Sally is a wonderful person, I say "things aren't always as they seem." *My* meaning—what I'm trying to get across—is that Sally is devious. But in my attempt to get this across, I'm taking advantage of the fact that "things aren't always as they seem" has an existing (linguistic) meaning. My meaning that Sally is devious is parasitic on my utterance's meaning that things aren't always as they seem. Meaning in the psychological sense is therefore parasitic on meaning in the linguistic sense.

1.2 Why Meaning #3 ≠ Meaning #2

Some philosophers and linguists have held that for:

(1) "snow is white"

to mean that snow is white is for it to be the case that, in uttering "snow is white," what *people* mean is that snow is white. This view, duly generalized, is that for a sentence S to have meaning M is for it to be the case that, in uttering S, people to mean M.[71]

This position is false. There are many sentences that have determinate meanings even though they've never been uttered before and, therefore, no one as of yet has ever meant *anything* by them. The sentence:

(2) "the cube root of three is Sir Lawrence Olivier's favorite irrational number between one and four"

has a determinate meaning, even though that sentence never probably has been uttered. Thus, meaning in the linguistic sense is not in *all* cases identical with meaning in the psychological sense.

But a stronger point is warranted. Let's say that P_1 and P_2 are the propositions meant by (1) and (2), respectively. In saying that (1) means P_1 and that (2) means P_2, we are not using the word "means" equivocally. Both occurrences of "means" in the last sentence denote the same relationship. Thus, the relationship that (1) bears to its meaning is the same as the relationship that (2) bears to *its* meaning. Given that, as we saw, (2)'s having P_2 for its meaning isn't identical with P_2's being what people mean in uttering (2), it follows that (1)'s having P_1 for its meaning isn't identical with P_1's being what people mean in uttering (1). Of course, what we just said about (1) and (2) can be said of any other sentence. So even if what people mean in uttering a given sentence happens to coincide with its literal meaning, what it is for sentence S to have proposition P for its literal meaning isn't for people to mean P in uttering S.

Psychological meaning presupposes linguistic meaning. What a person means when uttering a given sentence is a function of, among other things,

his beliefs as to what that sentence *already* means. You must believe that "snow is white" means *snow is white* if, intending to speak sincerely and literally, you say "snow is white" with the intention of getting it across that snow is white. If you think that "snow is white" means *grass is green,* you cannot, if your intention is to speak sincerely and literally, believe that "snow is white" means *snow is white.*

Of course, you could know full well what "snow is white" in fact means, but use that sentence to get across something that has nothing to do with the color of snow. Knowing what "snow is white" actually means, you might utter that sentence with the intention of getting it across that the government is controlling our thoughts with alpha waves. And, depending on the circumstances, that could be precisely what an utterance of that sentence would convey.

But whatever the message is that, in uttering a certain sentence, you wish to convey, you must believe that message to have *some* kind of relationship to the one meant by that sentence itself. Furthermore, if you are to *succeed* in saying what it is you wish to say, what you believe to be meant by the sentence you are using must be *right.* If, intending to speak sincerely and literally, you say "snow is white," thinking that it means *bananas are yellow,* you will fail to say what you wanted to say.

Thus, setting aside defective utterances, one cannot, in uttering a given sentence, mean *anything* by it unless one knows what it *already* means. So meaning in the *psychological* sense is parasitic on meaning in the linguistic sense, and the two kinds of meaning are therefore entirely distinct.

1.3 Why Meaning #3 ≠ Meaning #1

The sense in which "snow is white" means that snow is white isn't comparable to the sense in which smoke means fire. The fact that smoke means fire has nothing to do with conventions on the part of human beings. [72] But the fact that "snow is white" means what it does *is,* at least in part, a matter of convention. It's a matter of convention that "snow" doesn't refer to grass and, therefore, that "snow is white" doesn't mean that grass is white; it's a matter of convention that "white" doesn't mean *green* and, therefore,

that “snow is white” doesn’t mean that snow is green.

Although the whiteness of snow sometimes causes people to say “snow is white,” it doesn’t do so in the way that fire causes there to be smoke. Fire happens; smoke happens as a result. The presence of smoke doesn’t embody any *judgment* about anything. But, when caused by the whiteness of snow, utterances of “snow is white” *do* embody judgments of various kinds. People see or otherwise come to believe that snow is white; and, since they know the relevant linguistic rules, they know that, were they to say, “snow is white,” they judge that they’d be making a *correct* statement. Thus, utterances of “snow is white” embody judgments about the color of snow and about how, linguistic conventions being what they are, one can report the color of snow. Also, people don’t say everything that occurs to them. Before deciding to utter a given sentence, people typically make context-based judgments about the appropriateness of uttering that sentence. So various judgments—about snow, about language, and about human psychology—are involved in the causal connection between the whiteness of snow and a given gerson’s saying “snow is white.” There is thus a *normative* dimension to language use that is absent where purely natural, non-conventional cause-effect relations are concerned.

2.0 Sentences as proposition-isomorphs

The meaning of a true or false sentence is a ***proposition.*** Propositions are not themselves sentences. That is why different sentences (e.g., “schnee ist weiss” and “snow is white”) can express the same proposition.

Propositions, when true, are *truths.* Thus, propositions have existed as long as there have been truths; which means that they’ve existed as long as there has been anything and, consequently, that propositions are not creations of human creations.

Though distinct from the sentences that express them, propositions are structurally similar to them. Two otherwise dissimilar sentences can share the word “John.” “John loves Mary” and “Sally punched John” are two such sentences. The meanings of those sentences obviously have something in common corresponding to the fact that they share the word “John.” Since they share no other constituents, the thing meant by “John” must be capable

of moving *on its own* from sentence-meaning to sentence-meaning. This would not be the case if the thing meant by "John" in the proposition meant by "John loves Mary" were incapable of being disengaged from the things meant by "loves" and "Mary." It follows that propositions consist of discrete parts; it also follows that those discrete parts correspond to discrete parts of the sentences that express them. Taken together, these two points entail that sentences are structurally like the propositions they express.

2.1 Propositions as digital structures

Given that propositions consist of discrete, isolable entities, it follows that, like sentences and unlike visual perceptions and photographs, propositions are *digital* structures. The sentence "Sally punched Bob" has a unique decomposition into a certain "minimal units of significance," or "morphemes," these being "Sally," "Bob," etc. Given what we said in Section 2.0, it follows that something similar is true of the corresponding proposition. Sentences and propositions are *digital* structures, meaning that they have unique breakdowns into minimal significant units.

A visual perception of Sally punching Bob doesn't have a structure comparable to that of "Sally punched Bob" or any other sentence. Unlike sentences, perceptions don't have to decompose into minimal significant units. Visual perceptions, unlike sentences and propositions, therefore have a non-digital or *analogue* structures. Given that at least some thought involves the processing of perceptual information, it follows that thought

at least sometimes has a structure very different from language. (See Section 5.4 for further discussion of this.) The nature of propositions is discussed in Chapter 3.

3.0 The three branches of the philosophy of language: syntax, semantics, and pragmatics

The study of language is typically divided into three sub-disciplines—semantics, syntax, and pragmatics.[73] In addition to denoting a branch of linguistic study, each of these three words denotes dimension of language. So

"semantics" refers to a certain discipline and also to a feature of expressions, the same being true of the other two expressions.

3.1. Semantics

The discipline of semantics attempts to make it clear what our utterances literally mean. It has no interest in what is conveyed through suggestion or innuendo.

If a disappointed boss says to a substandard employee, "it might not be a bad idea for you to start thinking about finding a new position," the literal meaning of his utterance is quite innocuous. But the message that is being *sent* is not innocuous—that message is: *you're fired; you're a disgrace; go away;* etc. The utterance's *semantic* coincides with its literal (innocuous) meaning.

3.2 Pragmatics

Pragmatics studies the *use* of language. Sometimes language is used literally. Asked whether I'm over thirty years of age, I say "yes, I'm over thirty years of age." What *I* mean coincides with what my utterance literally means.

Language is often used non-literally. If, while addressing a pan-handler, I say "you've made a fine life for yourself," what I mean is the antitheses of what my utterance means. But usually the propositions literally meant by our utterances are neither opposed to, nor exhaustive of, the propositions we wish to affirm in producing those utterances. Asked whether there's a place to get food, I say "there's a McDonalds down the road." The proposition I'm affirming is: there is a nearby place to get food, the reason being that there's a McDonalds down the road. Thus, the proposition literally meant by my utterance is a only a part of what it is that I'm saying. Thus, what a sentence literally means is only one of many factors governing what it is used to mean. The discipline of pragmatics tries to identify the remaining factors.

3.3 Syntax

The disciplines of syntax studies the structures of the meanings of ***complex***

expressions. A complex expression is one that consists of other expressions. (Thus, "the man who ate my cookie" is a complex expression, since it consists of "man," "ate," etc., each of which is meaningful. By contrast, "red" is not a complex expression, since it doesn't have any meaningful proper parts.) The discipline of syntax tries to make it clear how the meanings of complex expressions depend on those of their parts.

Thus, the discipline of syntax doesn't study the meanings of complex expressions *per se.* It studies the *relationships* that such meanings have to those of their constituents. Consider the sentence "Sally hates Bob." The word "hates" ***occurs*** in that sentence. If that occurrence is replaced with an occurrence of "loves" or "is amused by," the resulting sentence has a very different meaning from the first. This shows that what "Sally hates Bob" means depends on what "hates" means.

Bearing this point in mind, consider the sentence "Larry loves Julie." Obviously this sentence doesn't mean the same thing as "Sally hates Bob." But the *relationship* borne by the meaning of "Sally hates Bob" to the meaning "hates" is identical with the relationship borne by the meaning of "Larry loves Julie" to that of "loves." Exactly similar points hold in connection with each of the remaining two constituents of each of those sentences.

The discipline of syntax studies the *relation* that the meanings of complex expressions bear to the meanings of the simple expressions composing them. Thus, syntax doesn't study the semantics (meanings) of complex expressions. It studies the *structures* of the semantics of complex expressions. Syntax studies semantic structure.

4.0 The need for the discipline of semantics

Even though we all know what:

(1) "John wants to catch a 20-pound striped bass"

means, we don't know what it is that we know in knowing this. Semanticists supply us with the missing metaknowledge. Consider the sentence:

(2) "John wants to punch Bob."

(2) attributes a certain property to John—that of wanting to punch Bob. Given that (1) and (2) are grammatically isomorphic, it's natural to assume that there exists some 20-pound striped bass x such that the proposition expressed by (1) is:

(3) John wants to catch x.

But this isn't the right analysis. There isn't some *one* fish such that, if the desire ascribed to John by (1) is to be gratified, John must catch *that* very fish. There is thus no fish x such that, if (1) is to be true, John must want to catch x.

The meaning of (1) is:

(1_R)[74] John wants it to be the case that: *there exists some fish x such that x is 20-pound striped bass and such that John catches x.*

Thus, (1) doesn't describe a relationship between John and some non-existent or quasi-existent fish. It affirms the existence of a relationship between John and a *proposition.* The proposition in question is one that, in English, is expressed by the sentence:

(4) "there exists some fish x such that x is a 20-pound striped bass and such that John catches x."

If John's fishing-trip is a success, that proposition will be true; otherwise it will be false. But that proposition exists either way.

But we still haven't solved the problem. In (4), the expression "some fish" occurs. Grammatically, that expression is a noun-phrase. But, unlike other noun-phrases, it doesn't refer to anything. ("Some fish" doesn't refer to some fish. There is no fish x such that "some fish" refers to x. That's why, given any particular fish F, if you say "some fish is wet, but x is not," what you are saying isn't self-contradictory.) So the problem we were trying to solve remains.

But to solve the problem, we need only reword (4). The needed rewording is this:

(5) The property of being a 20-pound striped bass that John catches is instantiated.

The property of being such a fish exists. So (1), which seemed to be about a

non-existent fish, is about an existent property. (5) says of that property that it's instantiated. Thus, a complete analysis of (1) is given by:

(1_{CA}[75]) John wants it to be the case that the property of being a 20-pound striped bass that John catches is instantiated.

So even though just about every English speaker understands (1), knowing what it is that one knows in understanding it isn't such a trivial thing.

4.1 Semantics needed to figure out what is literally meant and what is not

Despite everything just said, there is clearly a sense in which every English speaker knows what (1) means. What the semanticist is doing in connection with (1) isn't comparable to what you (who, we'll assume, speak Spanish) are doing in connection with it when you tell a monolingual Spanish speaking friend of yours what it means. The semanticist is needed to clarify the structure of the meaning that (1) is already known to have, but he isn't needed to *identify* that meaning. The semanticist isn't a translator. But there are many cases where the semanticist *is* needed to identify literal meaning. In fact, as paradoxical as it may sound, there are cases where he is needed to identify the meanings of sentences *that are perfectly well understood.*

First of all, we must distinguish what is literally meant by an utterance from what it is that the speaker wishes to convey. To give a trivial example: You and I are robbing a bank. I yell: "the cops are coming!" What I wish to convey is that we should hurry up. In this particular case, it's easy to distinguish what is literally meant from what is non-literally suggested, and semantics would therefore have no interest in it. But in other cases, it's exceedingly hard to do this, and it's with these other cases that semantics is concerned.

A story will help us move forward. Somebody who is wearing a ski-mask, and who I therefore don't recognize, deftly snatches my pocket from my wallet. As he's running off, I point at him and yell: "that man is a thief!" Let U_1 be this utterance.

Before moving on, let's take a moment to make it clear what U_1's literal meaning is. Somebody just stole my wallet. I don't know who that person is. But whoever it is, I am attributing a certain property to him. If that person has that property, I have spoken truly; if not, not. U_1 is correct if, and only if, the person referred to by "that man" has the property expressed by "is a thief." Thus, there is some individual x such that x has just stolen my wallet and such that what I've just said is true exactly if *x is a thief.* (The underlined part is U_1's literal meaning.)

The next day, my lovable office-mate Steve eats one of the cupcakes that was on my desk. I jokingly point at him and say: "that man is a thief." Let U_2 be *this* utterance. There is some x such that x just ate my cupcake and such that U_2 is true exactly if: x is a thief.

Unbeknownst to me, Steve is the pick-pocket, and there is some individual x, namely Steve, such that each of U_1 and U_2 is true if and only if x is a thief. Thus, U_1 and U_2 have the very same literal meanings.[76] But I don't know this, even though I speak English perfectly and, on each occasion, obviously understand perfectly well what it is that I'm saying.

How this is possible?[77] Our sense-perceptions describe things. My uttering U_1 was a response to my being given a visual description of Steve. That description was to the effect that:

(i) there is some man x such that x is wearing a ski-mask and such that x is running off into the distance.

But U_1's literal meaning is not that that effect. There is some man x such that x is a wearing a ski-mask (etc.), such that in uttering U_1 I was saying that:

(ii) x is a thief.

The meaning of U_1, being identical with (ii), is quite threadbare. But I grasped that threadbare meaning through my descriptively rich visual perception, whose content is given by (i).

My uttering U_2 was a response to my being given a different description

of Steve. That description was to the effect that:

(iii) there is some man x such that x is a portly amicable fellow who is sitting over in that chair.

But U_2's literal meaning is not that that effect. There is some man x such that x is a portly amicable fellow (etc.) such that in uttering U_2 I was saying that:

(ii) x is a thief.

Echoing what we said a moment ago, the meaning of U_2, being identical with (ii), is quite threadbare. But I *grasped* that threadbare meaning through my descriptively rich visual perception, whose content is given by (iii).

Because I grasped (ii) by way of different bodies of perceptual (descriptive) information, I didn't know, when uttering U_2, that what I was affirming was the same thing I was affirming in uttering U_1. Oftentimes, literal meaning is cloaked by the *pre*-semantic information through which it is grasped, and semanticists are needed to uncloak it.

4.2 Semantics needed to figure out what is literally meant and what is not (continued)

Fido is the smartest dog on the planet. I know this well, but some of my friends don't yet know this. I point to Fido and say: "that dog is very smart."

The proposition that it was my intention to affirm and communicate is indeed true. *For a dog,* Fido is indeed smart. Of course, Fido is vastly less intelligent than a human being, such as my friend Timmy, who is of mediocre intelligence. But if I say "Timmy is very smart," what I'm saying is *false.* Judging by the words I've used, the property I've attributed to Timmy is identical with the property I've attributed to Fido. Given that Timmy has that property to a vastly greater degree than Fido, it would seem to follow that, since "Fido is smart" is true, "Timmy is smart" must also be true. And yet "Timmy is smart" is false. How can this be?

Some deal with this by saying that "smart" is ambiguous, like the word "dumb." So "Timmy is smart" and "Fido is smart" have different meanings, like "Timmy is dumb [unintelligent]" and "Timmy is dumb [mute]."

This solution is pretty clearly false. A more plausible one is that the property of being smart *for a dog* is distinct from the property of being smart *for a human.* Fido has the first but not the second. And many humans have the second, but almost all of those lack the first.

A similar, possibly coincident, view is that "smart" is implicitly relational. When, for some object x, you say ‹x is smart›, you are saying that x is smart *relative to some benchmark,* the identity of which the context makes clear. So "Fido is smart" says that Fido is smart*er* than most dogs, which is true, and "Timmy is smart" says that Timmy is smarter than most human beings, which is false.

I published a paper[78] arguing that, for any degree-property phi, ‹x has phi› expresses a proposition of the form: the degree to which x has phi exceeds standard S, where S is some standard that, given the context, is clearly the relevant one. (A "degree property" is one that can be had to varying degrees.) But even if this is right, it doesn't follow that such judgments are the literal meanings of such sentences. And there is no independent evidence that ‹x is smart› has the syntactic properties of sentences that clearly *do* have for their literal meanings propositions of the just-described kind. This suggests that, so far as ‹x has phi› communicates such a proposition, it isn't because it semantically encodes it.

In any case, it not *obvious* what ‹x is smart› means or, in general, what ‹x has phi› means, where phi is any degree property. Thus, the literal meanings of such sentences are sufficiently recondite that the intervention of professional semanticists is needed to identify them.

5.0 Semantic rules

The English language assigns a certain meaning to the sound "that dog has rabies"; and given the spectacle of a rabid dog, it furnishes one with a sentence with which to describe what one sees. In general, the English language assigns meanings to sentences and sentences to meanings. This is true of all languages. A language is a systematic way of pairing off sentences with meanings. Any rule that assigns a meaning to an expression is known as a "semantic rule." Languages are sets of semantic rules.

5.1. An important subtlety

There are a couple of subtle but, in some contexts, important inaccuracies in what I just said. First of all, something isn't a sentence *until* a meaning has been assigned to it. In a world where there were no animate beings, but in which the forthcoming parenthetical ink deposit (1 + 1 = 2) was formed out of twigs, that twig deposit wouldn't be an expression of any kind. That twig deposit *would* be an expression *if* and *only* if it were endowed with meaning. This shows that something has to have meaning *in order* to be an expression. Thus, *expressions* aren't assigned meanings. They already have them and don't need to be assigned them. Therefore a semantic rule can't be defined as a rule that assigns a meaning to an expression.

5.2 How meaning is assigned to hitherto meaningless and, therefore non-linguistic, entities[79]

Thus, semantic rules assign meanings to non-expressions. But which non-expressions? A story will give us the answer.

You and I want to invent a code that only we two know. We both know a guy whose real is name is Larry. We decide that our code name for Larry is to be "Ichabod." So what's going on is that we're creating a semantic rule: one that to the effect that "Ichabod" is to pick out Larry. How exactly is this rule enacted?

In order to implement this rule, I say: Let's refer to Larry as "Ichabod." The burst of noise that I produce is a sentence-token. And what you *hear* is some *token* of "Ichabod"—you do not, since one could not, hear the name type itself.

When you hear this burst of noise, along with my proposal concerning our new name for Larry, you know that what I'm saying is to the effect that any other physical object that is similar in the relevant ways to this burst of noise is itself henceforth to refer to Larry. Thus, I am in effect proposing that all and only those bursts of noise that are similar, in the relevant respect, to this particular burst of noise are to refer to Larry. (By implication, I'm proposing the same thing *mutatis mutandis* to hold of all and only ink-

deposits that, given certain conventions, are paired off with such bursts of noise.) The thing that, according to my proposal, is henceforth to pick out Larry is the thing of which all and only such bursts of noises (etc.) are *instances.* That thing, like anything else of which there are instances, is a property. It is the property had in common by all and only bursts of noise (etc.) of the relevant type.

That property doesn't (yet) have a meaning; it isn't (yet) an expression. It's a property that existed, and was instantiated, before either or any instances meant anything. So the semantic rule that I'm proposing we adopt assigns a referent to a property that does not itself have a meaning. The same thing *mutatis mutandis* holds of any other semantic rule.

Thus semantic rules assign meanings to properties of physical objects—to morphological or acoustical properties (in other words, to properties that a things has in virtue of having a certain shape or sounding a certain way). A semantic rule is therefore something which assigns a meaning to a property, and language is a set of such rules.

(Technically, this is only an approximation to the truth. The relevant qualifications are found in Section 7.5.)

5.3 What are semantic rules?

Many believe that linguistic meaning is to be understood in function-theoretic terms—that, in other words, semantic rules are mathematical ***functions***.

Let us start by defining the word "function." Given any pair of whole numbers, the expression "+" assigns exactly one whole number to that pair. In general, a function is a rule that, given some class of objects, assigns no more than one object to any given member of that class.

Although the rule expressed by "plus" assigns the number 8 to the pair <4,4>, it doesn't assign that number to that pair in the way in which a person assigns a task to an underlying. In the former case, the word "assigns" has a psychological meaning; in the latter, it has a non-psychological, purely logical meaning. A related point is that the rule that assigns 8 to <4,4> isn't a social rule, and it therefore isn't something that can be obeyed or disobeyed.

According to the function-theoretic view, semantic rules are rules in the strictly logical sense; that is, they are mathematical functions. The semantic

rule for “Socrates” is a function that assigns a certain individual (Socrates) to that word (or to occurrences thereof). The semantic rule for “snow is white” is a function that assigns truth-conditions to that utterance (or to occurrences thereof). And so on.[80]

5.3.1 Why semantic rules are not functions

The just-described view is false. The rule denoted by “+” has always existed and always will. Of course, the *expression* “+” hasn’t always existed. But that’s irrelevant, since things pre-exist the expressions we use to denote them. “Socrates” is an Anglicization of the name with which Socrates referred to him. Since Socrates lived well before the English language came into existence, “Socrates” (the name, not the person) didn’t come into existence until well after its referent went out of it.

The semantic rule that assigns Socrates to ink deposits having certain shapes would exist even if the English language had never come into existence. Like the rule denoted by “+”, that rule has always existed, and always will. So has the rule that assigns Abraham Lincoln to such ink deposits. The mathematical function that assigns the proposition *snow is white* to ink deposits like the italicized one has always existed, as has the mathematical function that assigns the proposition *all horses weigh 18,000 lbs* to those same ink deposits.

But the English language hasn’t always existed. Since the English language *is* a set of semantic rules, those semantic rules haven’t always existed. Therefore, they haven’t always existed. They came into existence quite recently. Therefore, those rules aren’t mathematical functions.

Also, if the semantic rules of English *were* such functions, there would exist a language in which “Socrates” referred to Lincoln and in which “snow is white” meant *all horses weigh 18,000 lbs,* the reason being that the corresponding mathematical functions exist. But there is no such language. Of course, there *could* be such a language. And maybe there will be; maybe somebody will invent a code in which those things have those meanings. But right now they don’t. Such a language is merely possible and, therefore, doesn’t exist. Thus, semantic rules are not rules in the mathematical sense.

5.3.2 The Gricean approach

Understandably, many philosophers of language believe that semantic rules must be understood in psychological, not mathematical, terms. There are different versions of this view. I accept one version of it. But the version I accept bears little resemblance to the versions of it that are usually held, each of which is some variant of the view held by H.P. Grice.

According to Grice (1957), for expression E to have literal meaning M is for it to be the case that, when they utter E, M is what they mean. So "snow is white" has the proposition snow is white for its literal meaning because what people generally mean when they say "snow is white" is that snow is white. People generally mean snow is white in uttering that expression.

Literal meaning is to be understood in terms of speaker's meaning. That's the main idea. Neo-Griceans hold that, even though literal meaning cannot in all cases be identified with speaker's meaning, it is always, ultimately, to be understood in terms of it.

Wittgenstein (1958) advocated a version of this view. "Roughly speaking," he said, "meaning is use." Expressions mean what we use them to mean—they mean what *we* mean by them. Wittgenstein nowhere makes it clear what he means by the words "roughly speaking."

But it doesn't matter, since his statement isn't even roughly true. Literal meaning is isn't identical with speaker's meaning and isn't to be understood in terms of it. It's the other way around. We saw why in Section 1.3.

Also, Grice's theory fails to deal with the fact that the meaning of a subsentential expression isn't something that could be possibly be meant. "Of " has a meaning; so does "skip," "snorkel," "or," "gladly," etc. But whatever it is that "or" means, it cannot, at least not by itself, be what a *person* means. Obviously I can say "snorkel" and mean it—but only if I'm using it as an abbreviation for some whole sentence (e.g., "my favorite activity is to snorkel") and, therefore, to convey something other than its literal meaning.

5.3.3 More problems with the Gricean approach

If Grice were right, the meaning of the sentence:

(SF) "Smith is now living in France"

would be fixed by the intentions people have in using it.

But its meaning is *not* fixed by those intentions. It is fixed by the meanings of its parts ("Smith," "France," etc.), together with the way those expressions are ordered in that sentence. The semantic and syntactic rules of English being what they are, SF would have its current meaning even if it had never been used. So whatever the intentions of people using that sentence are, those intentions do nothing in the way of assigning it that meaning. In general, Grice's view is inherently incapable of accommodating the fact meaning is compositional.

5.3.4 How some Griceans deal with the problem just described

Some Griceans respond by saying that, although speaker-meaning doesn't *directly* fix sentence-meaning, it does so indirectly. In their view, it is because of what we mean by sentences of the form ‹....France...› that such sentences are to the effect that...*France*...and not to the effect that, for example,...*Germany*....[81]

In addition to being an abandonment of Grice's core idea, this move is a failure. Let P be the proposition meant by SF. So far as people utter SF with the intention of affirming P, it's because they believe (correctly, as it happens) that each of the expressions composing it already has a certain meaning.

It's irrelevant that how we use sentences of the form ‹...France...› *causally* determines what "France" means. There are many ways to *cause* meaning-shifts—many ways to *get* a given expression to have a certain meaning. But the question we're asking isn't "how did 'France' *acquire* its current meaning?", and is instead "whatever it is that 'France' means, what *is* it for it to have that meaning?" And it's no answer to this question to say that it may have acquired that meaning because of what, at some point in time, people meant by sentences of the form ‹...France...›.

5.3.5 Why Grice's theory is inconsistent with the normative nature of semantic rules

A billiard ball isn't right to move after being struck; it just does. The relevant scientific laws merely register that fact; they aren't normative—that is, they don't characterize it as good or bad. Unlike scientific laws, semantic rules *are* normative. If, intending to affirm that Smith is female, you say "Smith is male," you've done something wrong. If Grice is right, literal meaning is speaker's meaning. This means that, if Grice were right, speaker's meaning wouldn't be accountable to existing semantic rules. Since it is, Grice is wrong.

5.4 The psychological reality of semantic rules

One view as to the nature of semantic rules is that they are idealized descriptions of the activities of speakers. Proponents of this view seldom if ever identify the facts about the speaker-behavior of which semantic rules are supposedly descriptions. The most natural assumption is that they are idealized descriptions of what people *mean* when they speak and write (etc.). If this assumption is right, then, given the points just made, the view in question is wrong.

But even if this isn't what proponents of this view have in mind, their view is very clearly wrong. If semantic rules are just idealizations of speaker-behavior, then speaker-behavior must pre-exist the semantic rules embodied in it. But if that's the case, then the activity described by semantic rules isn't *guided* by them. An awareness of those rules is no part of what leads people to say the things they do. Those rules are psychologically inert. They have no "psychological reality." This view is held by Nathan Salmon (2007) and also by Scott Soames (2002).

This view is inconsistent with some obvious facts. I know that Smith is now living in France. Wanting to tell you this, I say: "Smith is now living in France." Why do I choose this particular sentence? Because, first of all, I know the relevant semantic rules (viz. that "Smith" refers to Smith, that "living" refers to a certain property) and, secondly, because I believe that, given these facts about semantics, the sentence in question is the right one to express my belief. A knowledge of semantics underlies my speech-act and, by obvious extensions of these points, all non-defective speech-acts.

Another problem with the Salmon-Soames view is that it's inconsistent

with the normative character of semantic rules. If semantic rules merely describe existing semantic activity, then that activity isn't answerable to semantic norms. It is; so the Salmon-Soames view is wrong.

5.5 Conceptual role semantics

A little while ago, we discussed Wittgenstein's claim that "meaning is use," i.e., that for an expression to have a given meaning is for it to be used in a certain way. This doctrine is incoherent in many ways. We've already discussed one of those ways; now we'll discuss some of the others.

Expressions have meanings. A meaningless burst of noise isn't an expression. If I cough or guffaw, the burst of noise I've produced doesn't have the sort of meaning had by *bona fide* expressions. It has, at most, meaning in irrelevant, purely evidential sense, e.g., the sense in which a cough may be evidence of a cold. Since anything that is a linguistic expression *ipso facto* has a meaning (in the relevant, linguistic sense), there are no expressions to be used *before* noises, ink-marks, etc., have been assigned meanings. So, since there can be no expression-use until after there is expression-meaning, it makes no sense at all to say that expression-use determines expression-meaning. "Meaningful expression" is a pleonasm.[82] So, contrary to what Wittgenstein said, meaning isn't use.

Of course, how a given expression is used may well assign it a *new* meaning. But there is all the difference in the world between saying:

(1) Expression E's having meaning M is *causally* determined by E's being used in such and such a manner,

and

(2) Expression E's having meaning M is *identical with* E's being used in such and such a manner.

An expression E's having meaning M cannot possibly be *constituted* by its being used in such and such a manner, since E isn't an expression and, therefore, isn't an expression to be used *until* it has a meaning.

According to a doctrine known as "conceptual role semantics" (CRS), whose exponents include Hartry Field (1977) and Robert Brandom (1994),

for an expression to have a given meaning is simply for it to be used in a certain way. But this isn't correct as we just saw.

CRS is incoherent for reasons other than the one just given. According to that doctrine, what a sentence means is determined by what people infer from it and what people infer it from. Whereas commonsense holds that one infers "an even number is less five" from "two is less than five" *because* the latter already has a given meaning, advocates of CRS say that, on the contrary, it's because people infer "an even number is less than five" and other similar statements from "two is less than five" that the latter has the meaning it has.

This is not a viable view. If we learn tomorrow that, contrary to what we previously thought, Aristotle wrote several plays, we'll infer "somebody who wrote several plays wrote the *Nichomachean Ethics*" from "Aristotle wrote the *Nichomachean Ethics*." But it doesn't follow that "Aristotle wrote the *Nichomachean Ethics*" would have undergone some change in its semantic *meaning.* Changes in what we believe affect what we infer from statements; but they don't categorically change the meanings of those statements. CRS entails that every inference is an analytic inference. Once it's learned that Aristotle wrote plays, it becomes, according to CRS, constitutive of the *meaning* of "Aristotle wrote the *Nichomachean Ethics*" that one can infer from it that a playwright wrote the *Nichomachean Ethics.* But surely the inference from "Aristotle wrote the *Nichomachean Ethics*" to "a playwright wrote the *Nichomachean Ethics*" isn't *analytic.*[83]

What we may infer from a sentence is answerable to its existing meaning. CRS says that a sentence's meaning is answerable to what we infer from it. If correct, that would have the consequence that one couldn't possibly draw a false inference from any sentence. Which, in its turn, would have the consequence that no sentence would mean anything. Which, since nothing meaningless is a sentence, would have the absurd consequence that there neither are, nor could be, sentences.

Consider the sentence "Bill plagiarized his first novel." If one knows that sentence to be true, one can make inferences about Bill's character, his past activities, his ambitions, his values, and so on, that one couldn't make if, other things being equal, one didn't know that sentence to be true. But it's only because of what the sentence *already* means that one can make those

inferences. Given any other sentence, the same thing *mutatis mutandis* is true of it. CRS says that what a sentence means is determined by what we infer from it. Since it's the other way around, as we've just seen, CRS is false.

6.0 What is literal meaning?

Where complex expressions are concerned, there is no limit to how much literal and understood meaning may diverge from each other. But where simple expressions are concerned, literal and understood meaning must coalesce. It makes no sense to suppose that people could be systematically wrong as to what "red" meant. If people *thought* that "red" meant what is in fact meant by "blue," then "red" *would* have that meaning. Systematic, widespread error is impossible where semantically simple expressions are concerned. This gives us a way of understanding what literal meaning is.

Even though there can be widespread, systematic misinterpretations of sentences, those misinterpretations do not arise as a result of people failing to know what the simple parts of sentences mean. They arise as a result of people not knowing how to put those meanings together. So *to the extent* that its meaning is fixed by the fact that it has the form ‹...Socrates...›, people (English speakers) *do* systematically understand "Socrates was more wise than Plato, but he was less sharp then Aristotle"; and *to the extent* that its meaning was determined by its having the form ‹...wise...›, people *do* understand that sentence; and so on. So far as that sentence *is* systematically misunderstood, it is because people are having trouble putting the meanings of its constituents together—it is because they're having trouble figuring out how those meanings ought to be put together.

What a simple expression literally means is determined by what it is that a sentence means by virtue of containing it. Since, where simple expressions are concerned, what people take literal meaning to be coincides with what it really is, a simple expression's literal meaning coincides with what it is that, in virtue of containing that expression, sentences are *taken* to mean. So a simple expression's literal meaning is given by a statement saying what it is that, by virtue of containing it, sentences are taken to mean; and a complex expression's literal meaning is a function, in the mathematical sense, of the meanings of its parts.

Here's an illustration. "Socrates" is a simple expression. So what people think it means must ultimately coincide with what it actually means. So it refers to Socrates only because people think it refers to Socrates and, therefore, only because people think that "Socrates is intelligent" attributes intelligence to Socrates and, in general, that ‹...Socrates...› attributes...x...to Socrates.

Of course, people don't always take utterances of ‹...Socrates... › to be attempts to *say* that Socrates has...x... It might be clear from the speaker's tone that what he *really* meant when in saying "Socrates was wise" was that Socrates was *not* wise. But *to the extent* that their belief that "Socrates" refers to Socrates is determinative of what people take the meanings of utterances of the form ‹...Socrates...› to be saying, what they take it to be saying is that Socrates has...x... That is what it is for them to take "Socrates" to refer to Socrates. And their taking "Socrates" to refer to Socrates *is* for "Socrates" to refer to Socrates, given that "Socrates," being a simple expression, has the semantics that people think it has. The same thing *mutatis mutandis* is true of every other simple expression.

Bearing these points in mind, let CE be any complex expression, and let $e_1...e_n$ be the simple expressions composing it. How people interpret CE may diverge from its literal meaning. But when this happens, it's because what it is taken to mean diverges from what, given what people believe its simple parts to literally mean, people are *disposed* to take it to mean.

Since what people take simple expressions to mean *is* what they mean, this is the same as saying the following. A divergence between

(i) a sentence's literal meaning

and

(ii) that sentence's understood meaning

is the same thing as a divergence between

(a) that sentence's literal meaning

and

(b) what it is that, given their (correct) beliefs as to what its simple parts literally mean, people are *disposed* to take that sentence's literal meaning to

be.

The literal meaning of a complex expression is a function of two things: (i) the meanings of its simple parts, and (ii) the order in which those parts are arranged. To say what literal meaning is in general, we need to say what it is for a simple expression to have a given literal meaning. Given the points just made, we can do this. Where simple expressions are concerned, literal and communicated meaning coincide (ultimately)—in other words, such expressions mean what people take them to mean. And where complex expressions are concerned, literal and communicated meaning may diverge, but literal meaning nonetheless coincides with what, given what the simple components of the expression in question literally mean, people are *disposed* to take it to mean.

7.0 Tokens vs. types: some preliminary terminological points

No word is identical with any utterance of it. My utterance of the word "snow" lasts for a fraction of a second. But the word "snow" *itself* endures.

Utterances and inscriptions of expressions are referred to as "expression-tokens" (or just "tokens"). So there are three *tokens* of some one word to the right of the upcoming colon: snow, snow, snow. The things being uttered or inscribed are referred to as "expression-*types*" (or just "types").

7.1 Two-dimensional semantics[84]

Some expressions appear to have a two-tiered semantic structure. For example, an occurrence of the pronoun "I" has a referent, this being the person who uttered it, and it picks out that referent in a certain way (i.e., by way of a certain concept).

Thus, the semantics of "I" is given by the rule that an utterance of "I" refers to a given person if and only if that person has the property of being the one that produced that utterance.

So, if I say "I am tired," the concept through which my utterance of "I" refers to me is the concept person who produced the utterance in question. This, then, is the concept through which reference is secured; it is, as we'll

henceforth put it, the mediating concept.[85] When Smith says "I am tired," his utterance of "I" refers to himself, not to me. But the mediating concept remains the same.

If somebody points to me and says "that guy is tall," the person picked out by "that guy" is me. But in this case, the mediating-concept is different. The semantics rule for "that guy" is given by the rule that, if "that guy" is uttered in a context where there is a unique, salient guy, that utterance refers to that individual.

"I" and "that guy" are context-sensitive expressions: what such an expression refers to depends in a systematic manner on facts about the context of utterance. (We'll soon refine this vague statement shortly.) Such expressions are known as "indexicals." Other examples of indexicals are "you," "he," "those animals," "this monkey," "tomorrow," "yesterday." Some indexicals are single words (e.g., "tomorrow," "he"); others consist of more than one (e.g., "that tall man"). The latter are known as "complex indexicals."

7.1.1 Demonstratives vs. indexicals

Some indexicals often cannot be successfully used without an accompanying demonstration on the speaker's part, the purpose of which is to eliminate any doubt as to what the intended referent is. If Jim and Larry are both equally salient in the context in question, an utterance of "that guy" won't single anyone out. But it will do so if an act of pointing accompanies it. Indexicals that fall into this category are known as "demonstratives." Not all indexicals are demonstratives. For example, "tomorrow," "now," and "here," aren't demonstratives. Given an utterance of "tomorrow," no gesture is needed to make it clear what the intended referent is. If I say "tomorrow I'm going hiking," it's clear what the referent of "tomorrow" is; no demonstrative act is necessary, and none could possibly do any good.

7.1.2 Indexicals (continued)

Indexicals have a "two-dimensional" semantic structure. Consider the expression "today." If I say it right now, that utterance will refer to April, 27, 2009, since that is today's date. If I say it tomorrow, that utterance will refer to April 28, 2009. But the *rule* that assigns April 27, 2009, to the first

utterance is identical with the *rule* that assigns April 28, 2009, to the second utterance. That rule is this: if "today" is uttered on a given day D, that utterance refers to D.

The meaning of an *utterance* "today" is the day it picks out. The meaning of the corresponding word-*type* is the rule just described. Given any indexical, the meaning of a *token* of that indexical is its referent; the meaning of the corresponding indexical-type is the rule that assigns that referent to that token.

7.2 Definite descriptions

In Chapter 6, reasons are given for thinking that definite descriptions are not devices of reference. But there are also reasons to think that they are devices of reference.[86] (It's very hard to believe that "the whole number that comes right after one" doesn't refer to the number two.)And in this section we will suppose them to be just that.

Definite descriptions, like indexicals, have a two-dimensional semantic structure. A given utterance of "the current U.S. President" refers to some individual. Right now such an utterance would refer to Barack Obama. A few years ago, such an utterance referred to Bill Clinton. But even though an utterance in 2009 of "the U.S. president" doesn't have the same referent as an utterance in 1999 of that same expression, both utterances are assigned their respective referents by the same semantic rule. The circumstances have changed—hence the change in referent—but the semantic rule has stayed the same.

That rule is: if, at time t, x uniquely has the property of being a U.S. president, then an utterance at t of "the U.S. President" refers to x.

7.2.1 Incomplete definite descriptions

Some definite descriptions appear to be "incomplete"—that is, they fail to pick out a single object. So, for example, "the bald guy" could pick out any one of many different people. But, like all expressions, definite descriptions are not used in a vacuum; and the context usually supplies the information needed to enrich the mediating concept enough to enable it to single out a single person. So if you and I are in a room and I say "the bald guy is

wealthy," it's clear that, so far as I making a determinate statement, what I mean is that the *contextually salient* bald guy is wealthy.

If phi is a property that obviously has no more than one instance (e.g., the property of being the whole number successor of one or of being the U.S. President), phi's sole instance is contextually salient by default.

7.3 The expressive limitations of indexical-free languages

Indexical-free languages are expressively impoverished; much of what there is to say can't be said in them. The reason is that much information is *perspectival,* and nothing perspectival can be stated in a language that doesn't contain indexicals. Suppose that I spoke a language that didn't contain any indexicals, but was otherwise just like English. In order to express what I believe concerning the weather in my area, I'd have to say "it's 60° in Richmond on Feb. 07, 2009." (And the "is" couldn't be taken as the *present tense* of the verb "to be," since, thus interpreted, it would be an indexical that picked out the time of utterance. The "is" would have to be stripped of any temporal meaning and, thus, be downgraded to an empty grammatical place-holder, like the "it" in "it's raining.") But that utterance wouldn't necessarily express the belief I wanted to express. My believing *it's 60° in Richmond on Feb. 07, 2009,* is different from my thinking *it's now 60° in Richmond,* even though I am in fact in Richmond at the time in question. I could have the one thought and not have the other. I could, after all, not know that I was in Richmond. And my believing either of those things is different from my thinking *it's now 60° here.* One could have any given one of those beliefs without having any of the others. Further, one could *rationally* have any given one of those beliefs without having any of the others. A person who is in Richmond at the time in question can *rationally* believe *it's now 60° here* while rejecting *it's 60° in Richmond on Feb. 07, 2009.* For the data at one's disposal may warrant the first judgment, but not the second.

So nothing perspectival—nothing that embodies any information relating to the speaker's perspective on the world—can be expressed without using an indexical. Nothing could be said about what's going on here, at *this* time; and there would be no me-thoughts (e.g., *I'm thirsty*) could be expressed; one

could only express their third-person counterparts (e.g., "JM is thirsty"). Thus, one could not, in an indexical-free language, express the thought *I am JM* or *I am that guy in the mirror.* And such a language would therefore be extremely impoverished.[87]

7.4 Tokens, types, and context-sensitivity

No sentence-*type* containing a context-sensitive component is either true or false; for no such sentence-type says anything. It is *tokens* of "I am now tired" that make statements; the corresponding sentence-type does not do so. What a token of "I am tired now" affirms depends on facts about the context of utterance. If Smith's the one who's speaking, and it's 3:00 P.M., such an utterance is true just in case Smith is tired at 3:00 P.M. If Jones is the speaker and it's 4:00 P.M., such an utterance is true just in case Jones is tired at 4:00 P.M.

Thus the semantic rule that assigns a proposition to such a token does so on the basis of the facts about the context of utterance. And the rule in question is clearly this: "If, at time t, person p tokens the sentencetype "I am tired," then the proposition thereby affirmed is true exactly if p is tired at p." The same thing *mutatis mutandis* is true of all context-sensitive expressions. So the rule that assigns a proposition to an utterance of "that man is married to that woman" is: if, in the context of utterance, x is a uniquely salient man and y is a uniquely salient woman, a token of the sentence-type "that man is married to that woman" affirms a proposition that is true if and only if x is married to y.

So context-sensitive sentence-types aren't true or false; they don't, in and of themselves, bear propositions. But they have an important semantic role: the identity of the proposition meant by a token of such a type depends on the identity of the type. A token of "I am tired" means one thing; a token of "you are tired," uttered by the same person at the same time, means something else; and that difference obviously stems from the fact that, because different sentence-types were tokened, the semantic rule that assigns a proposition to the one token is different from the rule that assigns a proposition to the other.

This can all be distilled into the following principle: where context-sensitive sentences are concerned, the meaning of a token is a proposition,

and the meaning of type is a *rule* that assigns a proposition to one of its tokens on the basis of facts about the context in which that token occurred.

Long story short: the meanings of sentence-tokens are propositions and the meanings of sentence-types are rules that assign propositions to their tokens, usually on the basis of facts about the context of tokening.

7.5 Ambiguity and context-sensitivity (revisited) and the typetoken distinction (revisited)

Consider the sentence-type "that person is a professor." Is that sentence true or false? No. Some tokens of it are true and some are false. This is because what a given token of that sentence says is a function of the circumstances. If I utter it while pointing at Bob, I'm attributing the property of being a professor to Bob. If I utter it while pointing to Sally, I'm attributing that property to Sally, not Bob. So what it is that I'm affirming in the one case is different from what it is that I'm affirming in the other case. But that isn't because the sentence "that person is a professor is ambiguous." That sentence is not ambiguous: it has just one meaning. But that meaning is not a proposition; it isn't something that is true or false. That meaning is a rule. That rule in its turn assigns meanings to tokens of that sentence. Those meanings are true or false; those meanings are propositions.

Remember that expression-types are what result when properties *per se,* as opposed to their instances, are assigned meanings. Of course, the rules that assign meanings to properties are semantic rules, since anything that assigns meaning to anything is *ipso facto* a semantic rule. So the sentence-*type* of which the following ink deposit—"that guy is a professor"—is a token is what results when some semantic rule assigns a meaning to some property. Let R_1 be the semantic rule in question. Let R_2 be the meaning that R_1 assigns to the just-mentioned property. R_2 is *itself* a semantic rule. But, whereas R_1 assigns a meaning (a rule) to some property, R_2 assigns meanings to instances of that property. So R_2 assigns meanings to particular instances of the morphology had by the following ink deposit ("that guy is a professor"). The meanings that R_2 assigns to those tokens are not themselves rules; they are propositions—they are things that are true or false.

There are thus very different sorts of semantic rules. There are those that assign meanings to properties, and there are those that assign meanings to instances of those properties. The meaning that was assigned to the property we were just discussing is itself a semantic rule. As we'll now see, *every* meaning that is assigned to a property (as opposed to a property-instance) is itself a semantic rule. Consider the following ink deposit: "Barack Obama." There is a semantic rule that assigns a meaning to the *property* of having a morphology similar to that ink deposit. The meaning assigned to that property is *itself* a semantic rule. *That* rule assigns a meaning to each instance of the property in question. It assigns Barack Obama, the person, to any such instance. Thus, any such instance is an expression that picks out Barack Obama.

In the previous section, we said that semantic rule is something which assigns a meaning to a property. That's correct, but incomplete. The right definition is this: a semantic rule is something which assigns a rule to a property that in its turn assigns meanings to instances of that property. Let us now move onto slightly less abstruse material.

7.6 Do any expressions have one-dimensional semantics?

The question arises: What about sentence-types that don't contain context-sensitive components? Where they are concerned, does token-meaning coincide with type-meaning?

First of all, in natural languages, there are no such sentence-types. Every sentence contains a tense-marker. And in virtue of containing a tense-marker, a sentence is such that the proposition expressed by any given one of its tokens is a function of (*inter alia*) when that token occurs. If you say "the U.S. economy is fairly stable," whether you are speaking the truth or not depends on when you say it.

There are some *apparent* counterexamples to this. For all intents and purposes, any two tokens of

(IA[88]) "the interior angles of a Euclidean triangle add up to 180°"

express the same proposition. But, from a strictly semantic perspective, IA *is*

context-sensitive, and tokens of it uttered at different times *don't* encode the same proposition. If it were believed that mathematical reality were as volatile as the stock-market, we would without hesitation regard IA as being in the same category as patently context-sensitive sentences such as:

(BH) "Bill's holdings add up to $180,000,000."

And we'd have no more temptation to regard the tense-marker in IA as inert than we'd have to regard its counterpart in BH as inert. What this shows is that, to the extent that the tense-marker in tokens of IA are doing nothing, it's only because, our beliefs about mathematics being what they are, we choose to see such utterances as expressing atemporal propositions.

Nonetheless, many semanticists hold that, at the level of literal meaning, IA is context-insensitive. They hold, in other words, that IA is what Quine (himself such a semanticist) refers to as an "eternal sentence." (S is an eternal sentence if there is some one proposition P such that any two of S's tokens express P.) If only for argument's sake, let's suppose that there are eternal sentences.

It's tempting to hold that an eternal sentence's meaning is identical with those of its tokens. After all, the distinction between type-meaning and token-meaning seems quite hollow *except* where context-sensitive expressions are concerned.

But this reasoning is spurious. The tense-marker on the occurrence of "add" in (IA) has the same semantics that it does in

(BH[89]) "Bill's various holdings add up to $180,000,000."

In BH the tense-marker is obviously doing real work in IA. That's why utterances of BH are true on Monday (before the market crashed) and false on Wednesday (after the market crashed). Therefore, different tokens of BH express different propositions. Therefore, the meaning of BH—the sentence-type—isn't some proposition, and is instead the rule:

(BHSR) If t is a token of BH that is uttered at time T, t is true iff, at T, Bill's various holdings add up to $180,000,000.

But, of course, for any time t, the meaning of a token of BH that is produced at t is a proposition that is true just in case.

(BHT) At t, Bill's various holdings add up to $180,000,000.

Given that occurrence of "add" in any given token of IA has the same literal meaning as its counterpart in any given token of BHT, it follows that, from a narrowly semantic perspective, IA is quite as context-sensitive as BHT. To be sure, there is obviously a sense in which IA is, whereas BHT is not, context-*in*sensitive.

But IA's context-insensitivity, we must conclude, is a thoroughly pragmatics-based affair. Obvious extensions of this reasoning show that *all* expression-types have *rules* that assign meanings (or referents) to their tokens and, therefore, that where any expression is concerned, type-meaning diverges from token-meaning. The meaning of the expression-type "that man" is: a token of "that man," uttered in context C, refers to x if, in C, x is uniquely a salient man. If, in a given context, Smith is such a man, then a token in that context of "that man" refers to Smith; if instead Jones is such a man, it refers to Jones; etc. The same is true of "the current President." In one context (the year 1992), tokens of that expression refer to Bill Clinton; in a different context (the year 2009) it refers to Barack Obama.

Some would be tempted to say that, because they have fixed referents, proper names (e.g., "Bill Clinton") don't have a two-dimensional semantic structure. This would be a mistake. Expression-tokens are physical entities—bursts of noise, deposits of ink, etc. Expression-tokens are therefore perceptible entities and are thus capable of transmitting information. Expression-types, on the other hand, are abstract entities and are thus inherently unsuited to be vehicles of communication. Let us develop these points.

Expression-types are properties. Why are they properties? Tokens of a type are instances of it. Anything of which there are instances is *ipso facto* a property. So expression-types, unlike expression-tokens, are properties and are therefore non-spatiotemporal entities. And, as we just noted, it makes no sense to suppose a nonspatiotemporal entity could mediate information or, therefore, could in any significant sense be a symbol. Also, given the profound metaphysical differences between tokens and types, it would be theoretical arbitrariness of the worst kind to suppose that types could discharge the same semantic functions as their tokens. We must therefore assume that, whereas tokens of "Bill Clinton" refer to Bill Clinton, the corresponding type does not refer to Bill Clinton; and we must also assume

that “Bill Clinton,” the expression-type, has for its meaning a (constant) function that assigns Bill Clinton to any given one of its tokens. In general, proper names, no less than indexicals and definite descriptions have two-dimensional semantics.

The semantic rule corresponding to a proper name is a constant function, whereas the semantic rule corresponding to an indexical or definite description is not a constant function. All utterances of “Bill Clinton” refer to Bill Clinton, but not every utterance of “the current U.S. president,” or of “that guy over there,” so refer. This has encouraged the erroneous view that “Bill Clinton” itself, the expression-type, refers to Bill Clinton.[90]

7.7 Ambiguity and context-sensitivity (re-revisited) and the type-token distinction (re-revisited)

Remember that expression-types are what result when properties *per se,* as opposed to their instances, are assigned meanings. Of course, the rules that assign meanings to properties are semantic rules, since anything that assigns meaning to anything is *ipso facto* a semantic rule. So the sentence-*type* of which the following ink deposit—“that guy is a professor”—is a token is what results when some semantic rule assigns a meaning to some property. Let R_1 be the semantic rule in question. Let R_2 be the meaning that R_1 assigns to the just mentioned property. R_2 is *itself* a semantic rule. But, whereas R_1 assigns a meaning (a rule) to some property, R_2 assigns meanings to instances of that property. So R_2 assigns meanings to particular instances of the morphology had by the ink deposit (“that guy is a professor”). The meanings that R_2 assigns to those tokens are not themselves rules; they are propositions—they are things that are true or false.

There are thus very different sorts of semantic rules. There are those that assign meanings to properties, and there are those that assign meanings to instances of those properties. Consider the following ink deposit: “Barack Obama.” There is a semantic rule that assigns a meaning to the *property* of having a morphology similar to that ink deposit. Let PBO be that property. The meaning assigned to PBO is *itself* a semantic rule. In other words, the

meaning of PBO is given by the statement that:

B_1: Any given token of PBO refers to to Barack Obama.

Given some specific token t of "Barack Obama", the semantic rule for t is:

B_2: t refers to Barack Obama.

B_1 and B_2 are very different rules. Unlike B_2, B_1 doesn't say anything about any *specific* token of "Barack Obama" or any other expression. The need for a two-dimensionalist approach to semantics is embedded in the very concept of what an expression is.

8.0 Logical form[91]

Some statements that, given their grammatical forms, appear to be about objects are in fact about properties. We saw this in Chapter 1, and we'll see it again in Chapter 7 when we discuss Frege's ground-breaking insights into arithmetical statements (e.g. "2 + 3 = 5"). Frege was the first to see clearly that logical and grammatical may diverge—he was the first to grasp the very idea of such a divergence. This insight of his is embodied in his statement that the sentence

(WM) "whales are mammals"

isn't about whales. As paradoxical as it may seem, he was right. WM says that *if* an object is a whale, *then* it's a mammal. But there is no specific object x such that WM says that x is a mammal. *A fortiori* there is no specific whale x such that WM says that x is a mammal; and for any number n, no matter how high, there are no whales $x_1....x_n$ such that WM says that x_i $(1 \leq i \leq n)$ is a whale. So just as Frege said, WM isn't about whales.

WM makes a statement, not about whales, but about the property of being a whale. It says that

(WM*) the property of being a whale has the property of being instantiated only by mammals.

WM* perspicuously represents WM's meaning. In other words, WM* represents WM's logical form. WM*'s grammatical form diverges from

WM's grammatical form.

The reason is that WM's logical and grammatical forms pull apart is that WM contains a ***quantifier.*** (Examples of quantifiers are "all birds," "no man," "most whales," "three birds," and "some individuals." We'll define the term "quantifier" in a moment.) Given *any* statement of English, or any other non-artificial language, that contains quantifiers, the logical and grammatical forms of that sentence diverge. A quantifier is an expression having the property that, if a sentence contains it, that sentence is *ipso facto* to the effect that the ***extension*** of one property has a certain degree of overlap, ranging anywhere from no overlap to total overlap, with the extension of some other property. "All birds have beaks" says that the extension of the property of being a bird is a subset of the extension of the property of having a beak. "Only birds have beaks" says (falsely) that the extension of the property of having a beak is a subset of the extension of the property of being a bird.

In virtue of a containing a quantifier, a sentence about properties, not about specific objects. Of course, a quantified (quantifier-containing) statement can *also* be about individuals. But it isn't in virtue of containing a quantifier that a sentence concerns individuals; and it *is* in virtue of containing a quantifier that a sentence concerns properties.

In natural language, quantified sentences are pseudo-objectual statements. They appear to be about objects, but are really about properties. That is why they must be reparsed if their logical forms are to be exposed. It immediately follows that a statement's grammatical form may diverge from its logical form. Frege discovered both facts and, therewith, created modern analytic philosophy.

8.1 Contextual definition

The fact that grammatical and logical form sometimes diverge is related to the fact some expressions are to be defined contextually. To define an expression contextually is to how, in virtue of containing it, a sentence's meaning is affected. So, for example, "someone" doesn't refer to anyone. It doesn't refer to John or Sally or Jane. That is why, given any proper name N, the statement

(NS) ‹N doesn't snore ›

is ***compatible*** with[92]

(SS) “someone snores.”

Thus, the meaning of “someone” isn’t given by some rule that pairs it off with this or that individual. In other words, there is no individual N such that the semantic rule for “someone” is:

(WRS[93])‹Someone has psi› is true just in case N has psi.

Rather, the meaning of “someone” is given by the statement that:

(RS) ‹Someone has psi› is true just in case the property of being a psi is instantiated.[94]

Thus, the logical form SS is clearly displayed by the sentence:

(PS[95]) “the property of being a snorer is instantiated.”

PS’s logical form therefore coincides with its grammatical form. Since SS has a different grammatical form from PS, it follows that SS’s logical form diverges from its logical form. The divergence between SS’s grammatical and logical forms is obviously a consequence of the fact that “someone” must be defined contextually.

8.2 Contextual definition: its scope and limits

Nonetheless, it would be an overstatement to say that *whenever* a sentence contains an expression that must be defined contextually, it’s grammatical form pulls apart from its logical form. As we’re about to see, *every* expression is to be defined contextually. Obviously grammatical form doesn’t *always* pull apart from logical form. Therefore, it isn’t *always* the case that, in virtue of containing an expression that must be defined contextually, a sentence’s logical and grammatical forms diverge.

How can it be said that all definitions are contextual definitions? Aren’t there also ***denotative definitions?*** (A denotative definition of an expression E says what E means by saying what it denotes.) Isn’t “Socrates” defined denotatively? Yes, it is. There is some object x such that one says what “Socrates” means if, and only if, one says that “Socrates“ refers to x.

But in saying of some object x that “Socrates” picks out x, one is saying

that, in virtue of having the form ‹Socrates has psi›, a sentence S is to the effect that x has psi. “Socrates” refers to Socrates because, the remaining semantic rules of English being what they are, if you wish to attribute the property of being wise to Socrates, you can do so by saying “Socrates is wise”—because, in general, for any property psi, if you wish to attribute psi to Socrates you can do so saying ‹Socrates has psi.› If, in saying “Socrates was wise,” you were really saying that it was Aristotle, not Socrates, who was wise, then “Socrates” wouldn’t refer to Socrates, at least not in that context. In general, to say that E refers to O is to make a statement about effect a sentence’s containing E has on its truth-conditions. More precisely, it is to say that, in virtue of having the form ‹E has psi›, a sentence attributes psi to O, for any property psi.

But the logical forms of sentences of the form ‹E has psi› don’t necessarily pull apart from their logical forms. Since there is some object x (namely, Socrates) such that ‘Socrates is tall’ says that x is tall, and since ‹x is tall› has the same form as “Socrates is tall,” the fact that “Socrates” is to be defined contextually does *not* entail that logical and grammatical form ever diverge.

So it is only because *certain* expressions are to be defined contextually that such divergences occur. But which expressions? *Those that cannot also be defined denotatively.* Whenever an expression can be defined denotatively, a sentence’s logical form will not, by virtue of that sentence’s containing that expression, diverge from its logical form. (But, of course, that sentence’s logical form may diverge from its logical form for some other reason. Thus, the logical form of “Socrates saw someone” diverges from its grammatical form; but the reason for this is that it contains the word “someone.” The occurrence of “Socrates” in that sentence isn’t what induces that divergence.)

Let E be an arbitrary expression that can be defined denotatively. In other words, suppose there to be some object x such that E is defined by saying that it refers to x. The logical form of ‹E has psi› is: *x has psi.* In general, *to the extent* that the logical form a sentence containing E is determined by its containing that expression, that sentence’s logical and grammatical forms coincide. But, since such a sentence’s logical form isn’t determined *only* by its containing E, and is also a function of the semantics of the other expressions occurring it, its grammatical form may still diverge from its

logical form. After all such a sentence may contain an occurrence of "someone" or "nobody" or some other expression that induces such a divergence.

8.3 Frege's generalization of the concept of a function[96]

As we'll see in Chapter 7, what made it possible for Frege to revolutionize logic was his insight that grammatical and logical form diverge; and what made the latter insight possible was his generalization of the concept of a mathematical ***function.***

A mathematical function is a rule that assigns no more than one object to each object falling in a given class. (So "+1" can be thought of as expressing a function or rule that assigns 2 to 1, 3 to 2, etc.)

According to Frege, the occurrence of "snores" in

(PS) "Plato snores"

is best represented as the open sentence ‹x snores›; and that open sentence is best thought of as expressing a function that assigns the truth-value *true* to each snorer and the truth-value *false* to each thing that doesn't snore. (For brevity's sake, I'll henceforth use the words "truth" and "falsehood" instead of, respectively, "the truth-value *true*" and "the truth-value *false.*")

Here's the idea. A true sentence results if the variable in ‹x is even› is replaced with an expression denoting an even number and false if it's replaced with a number that doesn't denote such a number. ("2 is even" is true and "3 is even" is false.) So we can think of ‹x is even› as assigning truth to two, four, etc., and falsity to one, three, etc. For similar reasons, we can see ‹x snores› as assigning truth to snorer Bob and falsehood to non-snorer Wilma, and so on.

Frege sees the occurrence of "is taller than" in

(BTM) "Bill is taller than Mary"

as being identical with the open sentence ‹x is taller than y›, and he sees that open sentences as expressing a function that assigns truth-values to *ordered pairs* of objects. So ‹x is taller than y› assigns truth to the ordered pair <x, y>

if x is taller than y and otherwise assigns falsehood to that pair.

Frege treats:

(BJMP) “Bill is standing between Mary and John”

as comprising the open-sentence ‹x is standing between y and z›; and he sees that open sentence as expression a function that assigns truth to ordered triples of objects—as assigning that truth value to <Bill, Mary, John> just in case Bill is standing between Mary and John.

In this way, Frege was able to assign a *single* form to all ***atomic*** sentences. An atomic sentence is one that, unlike “Bill is tall and Sally is smart,” doesn’t consist of other sentences and that, unlike “someone snores,” doesn’t contain quantifiers. Non-atomic sentences are ***molecular.*** A molecular sentence is one that either consists of other sentences or contains a quantifier. Frege was able to show that, ultimately, all atomic sentences have the form ‹O has psi. › Contrary to first appearances, BTM has the form: *the ordered pair <Bill, Mary> has psi,* where psi is the property had by such a pair just in case its first member is taller than its second.

Pre-Fregean logicians saw each of PS, BTM, and BJMP as having a different form from each of the other two, and this made it impossible for them to do anything meaningful in the way of formalizing inferences involving atomic sentences. But Frege didn’t have this problem, since he, unlike them, wasn’t made blind by grammatical surface structure to the underlying structural similarities.

Chapter 5

Do We Think in Words?

1.0 Two different ways that knowledge of a language might enhance intelligence

It is obvious that knowledge of a language makes one more intelligent in at least some respects. Consider the thought that authoritarian governments tend to be less responsive to the needs of their constituents than democratic

governments. Could one grasp that thought without knowing some language or other? We have, I think, at least *some* inclination to say "no." And, more generally, we are reluctant to regard creatures that cannot interact with reality through a language, or in any case through some sort of symbolic medium or other, as being capable of grasping, let alone delineating the consequences of, abstract principles or therefore of thinking about reality in a *discursive* manner.

My own view, as we'll see shortly, is that this is false—that whatever one can grasp *through* language, one must be able to grasp *independently* of language. But this, I grant, is a counterintuitive view; and, speaking personally, it was only after trying to defend the negation of that view, and finding that it was impossible to do so, that I came to accept it. Still, our untutored intuitions do, I think, incline us to hold that, but for knowledge of a language (of some kind or other), a creature would be incapable of thinking discursively about the world.

But untutored intuitions are not always probative. And sometimes they are inconsistent with one another, this being one of those cases. For, while we are reluctant to impute discursive thought to the language-less, we *also* strongly feel that we sometimes have thoughts of considerable sophistication that we don't know how to put into words; and, if that's right (as I personally have no doubt that it is), then it seems to follow that not all discursive thought is language-dependent.

The purpose of this chapter is to provide a brisk but relatively rigorous answer to the question: Do we think in language?

2.0 Two different versions of the thesis that we think in language

Do we think in language? No. But, before we can say why, there are two very different versions of the view that we think in language.

Version #1: English speakers think in English; Albanian speakers think in Albanian; etc. In general, people think in the *natural* languages that they learn.[97] (A "natural language" is one that organically arises through human interactions. So English, Spanish, Arabic, Chinese, etc. are natural languages.

A non-natural or "artificial" language would be one that people invent—for example, computer-programmers invent languages to serve their special needs, and so do mathematicians and philosophers. Philosophers invent special notations because natural languages often obscure important features of the logical structure of the meanings that their sentences express.)

Version #2: We think in an innately known code, a "language of thought." So although we think in sentences, those sentences don't belong to English or Spanish (etc.) but rather to some system of symbolism that is hard-wired into us from birth.[98]

3.0 Why version #1 is wrong

Do English speakers think in English, Spanish speakers in Spanish, etc.? No. (We will find that some of the arguments against Version #1 also apply to Version #2.)

3.1 First reason

We can obviously have thoughts that we are incapable of putting into words. This shows that we grasp those thoughts in some way *other* than by grasping the English sentences that express them. Also, there are occasions where, although we *can* put into words what we are thinking, doing so requires great effort: we have to think long and hard about which words to use. This shows that thinking is often a *pre*requisite to using language, which in turn shows that at least some thinking is non-linguistic.

Astonishingly, many authors (e.g., Ludwig Wittgenstein, Simon Blackburn, John McDowell) *deny* that we know anything that we can't readily articulate. "If you can't say it, you don't know it," they say. This presupposes that what we believe is expressed *only* in the words we utter. But that's not the case. A given person's beliefs can be expressed in many ways, and many of them don't depend on that person's being able to use words. A person's knowledge of the differences between x and y may be expressed in that person's drawings of x and y; or it may be expressed in that person's movements while he is in the vicinity of x and y; or it may be expressed in some other non-linguistic manner.

The way that authors such as Blackburn, McDowell, Wittgenstein respond to this is to say that no *bona fide* knowledge is involved in my painting things accurately—that what is involved is a kind of pseudo-knowledge. This is obviously false. It also overlooks the fact that, but for this "pseudo-knowledge," we couldn't know anything about language, since it is only through our ability to differentiate noises, ink-marks, etc., on the basis of their perceptible properties that we can learn, or use, language.

3.2 Second reason

Some kinds of information (e.g., visual information) seem inherently incapable of being fully expressed by sentences of English or of any other natural language. Also, sentences have very different structures from pictures, even when the sentences and the pictures overlap in content. A visual perception of a man wearing a leather jacket has a very different structure from any sentence that describes that experience. For example, such a sentence decomposes into a finite number of distinct parts (words), whereas the sense-perception does not so decompose. (Sentences are ***digital*** structures, meaning that they consist of discrete entities. Sense-perceptions are ***analogue*** structures, meaning that they don't.)

Also, any sentence must be interpreted to be understood. To understand the sentence "there is a man wearing a leather jacket over there," you must know the relevant semantic rule (i.e., you must know the rules that assign meaning to that sentence). But you don't have to interpret a visual perception, at least not in the same way. Perceptions are "self-interpreting" in a way that sentences are not.[99]

3.3 Third reason

In order to learn a language, one must already be able to think. To learn English is to learn that certain sounds have certain meanings; and in order to be able to do this, you already have to be able to synthesize information—to make judgments, on the basis of experience, as to what sounds have what meanings. So knowledge of a language *pre*supposes the ability to think.

3.4 Fourth reason

Different sentences can have the same meaning, and a single sentence can have multiple meanings.

What you think when you think that Bob loves Mary is identical with what you think when you think that Mary is loved by Bob. But the sentence "Bob loves Mary" is different from the sentence "Mary is loved by Bob," even though they mean the same thing. So if you thought in English sentences, then the belief you'd have by virtue of accepting the sentence "Bob loves Mary" would be totally different from the belief you'd have by virtue of accepting the sentence "Mary is loved by Bob." But the belief you have in virtue of accepting the one sentence is identical with the belief you have in virtue of accepting the other.

Also, if you thought in language, you couldn't disambiguate ambiguous sentences. If your thinking *Bob is dumb (i.e., unintelligent)* were identical with there being an occurrence in your mind of the sentence "Bob is dumb," then your thinking *Bob is dumb (i.e., he can't speak)* would be identical with there being an occurrence in your mind of that same sentence going off in your head, and those thoughts would therefore be identical. But they are not identical. So your thinking *Bob is dumb (i.e., unintelligent)* is not identical with your thinking "Bob is dumb."[100]

3.5 Fifth reason (similar to reason #4)

People who speak different languages can have the same thoughts. Monolingual Spanish speakers believe that 1 + 1 = 2, and so do monolingual English speakers. But English and Spanish speakers use different sentences to express that truth, showing that thinking that truth cannot be identical with thinking a sentence of English or Spanish or, by obvious extensions of this line of thought, any other language.

3.6 Sixth reason

Understanding sentences involves thought. You must be able to think to understand what is meant by "if snow is cold, then snow is not hot." This is because, in order to understand that sentence, you must be able to pair it off with the right meaning, and your pairing that sentence off with the right meaning is obviously different from an image of that sentence flashing

through your mind. If I hear a sentence of a language that I don't understand, it may pass through my mind later (i.e., I'll remember the sounds later, just as songs that one has heard often flash through one's consciousness), but I won't know what it means. So understanding a sentence is different from its flashing through your mind. But if your thinking *if snow is cold, then snow is not hot* were identical with your thinking "if snow is cold, then snow is not hot," then your having that thought *would* be identical with that sentence's flashing through your mind. Since it isn't, we don't think in natural language.

4.0 Why version #2 is false

Do we think in some innate code? No. (Some of the arguments that we used against Version #1 obviously apply to Version #2. So we won't repeat those.)

4.1 First reason

We often think in images (visual and auditory), and the contents of these images seem to have a structure that is fundamentally different from the structure of any sentence. Sentences are "digital" structures, meaning that they decompose into a finite number of distinct parts. Sense-perceptions are non-digital; they are "analogue" structures, meaning that they *don't* decompose into discrete parts and are thus characterized by a kind of seamlessness that sharply distinguishes them from sentences.

4.2 Second reason

Knowing a language involves understanding the expressions belonging to it. Understanding such expressions involves knowing the rules that assign meaning to them. One can't understand any language, including an innately known one, without knowing such rules. But if one's thought is entirely dependent on such a language, then one can't know those rules, or anything else for that matter, without understanding those expressions. Therefore it is viciously circular to hold that we think in an innately known code.

4.2.1 How advocates of the view in question deal with this argument

Advocates of the view that we think in an innate code deal with this problem

by saying that, strictly speaking, we *don't* understand the sentences of our innate code. We are merely "built to conform" to those sentences. In other words, although we don't understand those sentences, we are so neurologically structured that they cause us to act in ways that are appropriate to the meanings of those sentences. So if the sentence of my innate code that means *fire is hot* goes off in my head, that fact will *cause* me to go near fire if I'm cold and to stay away from it if I'm hot (or some such).

4.2.2 Why this move doesn't work

If a sentence isn't understood by so and so, then from so and so's viewpoint it is not functioning *as* a sentence. If, as advocates of the viewpoint just presented maintain, nobody understands the sentences that mediate their thoughts, then those sentences aren't really functioning *as* sentences and, for all intents and purposes, they cease to be sentences.

Here's an analogy. Suppose that you have a book, but that, instead of reading it, you use it as a weapon—you throw it at your roommate, of whose antics you are growing increasingly tired. Even though the thing you are using as a weapon is a book, it is not in this context functioning *as* a book and, for all intents and purposes, is not a book. Similarly, if nobody understands the sentences that are, supposedly, identical with their own thoughts, then those sentences aren't really functioning as sentences anymore —they're functioning more like the copy of *War and Peace* that, instead of being read, is being hurled at one's roommate. In any case, to say that our thoughts are sentences that we never understand is to take a very artificial and implausible view.

So we may tentatively conclude that we *don't* think in language, even though, undeniably, knowledge of language enormously enhances at least some of our cognitive capabilities.

5.0 Given that thoughts aren't identical with sentence-tokens, why is it that knowledge of language enhances cognitive ability and why does it feel as though we think in words?

On your computer desktop, there are various "icons." By manipulating these icons, you can perform various operations—you can close a document, open your email account, etc. But these icon-manipulations are but representations of changes in patterns of electrical activity to which they bear only an extremely schematic resemblance.

When I'm thinking—when, for example, I'm trying to solve a problem (especially one connected to my work)—my thoughts often take the form of a voice. I have mental images of spoken words. This also happens when I'm reading. Sometimes it happens that I read one book for a while (e.g., one by Frege) and then start reading another (e.g., one by comedian Dave Barry), and I'm startled as to why *Frege* would say such things. Then I realize that I'm no longer reading Frege and that I should therefore "switch voices."

And, of course, people think with the help of visual imagery. Many a student of calculus, or even of more advanced branches of mathematics, relies on mental images. I know I do.[101] But the images that go through my head when solving some problem aren't *identical* with the ratiocinative activities in which my problem-solving thought consists. Taken by themselves, those images are quite feeble. They don't specify the exact shape of the object I'm thinking about. I'll use a five-sided figure to stand for a ten-sided figure, since I can't visualize the latter. But it doesn't matter, provided that I can find *some* mental icon that will stand for the relevant concept. The icon does not itself have to depict the concept or object it represents. It need only be sufficiently differentiated that I won't confuse it with the image I am using to stand for some other concept of mine.

But even if I were more visually adept and, for that reason, could produce more exact visual surrogates for the concepts I'm thinking about, my ratiocinative activity couldn't possibly be *identical* with any sequence of images. For any given sequence of images, a creature that could experience those images wouldn't necessarily have any of the concepts that they represent in my thought processes. No matter how *conceptually* impoverished a creature is, little or nothing follows as to what sorts of images can go off in its head.

But everything constitutive of *consciousness* seems to be image-like or, at the very least, to share with images the property of being

phenomenologically pregnant.[102] Wittgenstein (1958) concluded from this that thinking isn't a private psychological act. For reasons that we'll discuss in the next section, he inferred that thinking consists of engaging in overt behaviors that involve symbols belonging to some public language.[103]

Wittgenstein's position is rank absurdity. He's right, of course, that tickles, itches, etc., aren't thoughts.[104] But what follows isn't that thoughts aren't mental entities. What follows is that thoughts, unlike itches and tickles, aren't phenomenologically pregnant.

Incidentally, the mistake Wittgenstein is making here is very similar to a mistake that David Hume makes in connection with the nature of personal identity. David Hume argues that there is nothing to one's mind other than the conscious events that populate it.[105] His argument is that, when introspects, he only encounters various tickles, pains, and other sensations. This argument assumes that everything there is to know about one's mind is to be known in the same quasi-perceptual, phenomenology-drenched way in which we know of our sensations. But this assumption is blatantly false. I know that I believe that 2 + 2 = 4, and I know it in a relatively direct way. But it isn't because I experience some *sensation* or mental image or other phenomenally pregnant entity that I know this.

Wittgenstein's view that thinking consists of manipulating expressions is inconsistent with the fact that, unless one can think, one cannot use expressions meaningfully. If I just bark out sounds, having no idea what they mean, I'm not in any real sense using language. For my noise-making to be speech, it must be guided by a knowledge of what those noises mean.

Wittgenstein is right that the sensations and images that pass through one's mind when one is having a given thought do not *constitute* that thought. But he's wrong to conclude that thinking isn't a psychological process. (As previously stated, what he should have concluded is that not all psychological activity consists of images and sensations.) He's also wrong to conclude that those images categorically have no significant cognitive function. Mental imagery facilitates thought. Thanks to it, we have thoughts that we otherwise simply couldn't have.

Given this data, I would suggest the following. Though not themselves

constitutive of thought, mental images allow us to manage thought-processes that we otherwise couldn't manage. Just as one can manipulate otherwise uncontrollable patterns of electrical activity by manipulating the icons on one's computer desktop, so one can manipulate otherwise uncontrollable patterns of ratiocinative activity by manipulating cognitive imagery.

There is another point to make. Consciousness is a place where there are different streams of cognitive activity—where otherwise mutually isolated streams of cognitive activity can meet. The processes that create an olfactory or auditory perception are isolated from those involved in reading music or understanding facial expressions. But, in consciousness, sense-perceptions meet with judgments and with intentions to act, and so on. Consciousness has the effect, and maybe the purpose, of integrating otherwise mutually isolated streams of cognitive activity and, therefore, of pooling otherwise discrete bodies of knowledge. Mental imagery improves thought because it enables otherwise compartmentalized ratiocinative activities to borrow information from other similar activities from which they would otherwise be sealed off.

By virtue of knowing a language, one automatically has at one's disposal a rich and easily operated system of icons by means of which one can manage one's cognitive activities in the way just described. This, I would suggest, is why knowledge of a language does so much for thought.

6.0 Wittgenstein on language, thought, and the relationship between the two

Ludwig Wittgenstein (1889–1951) put forth several provocative contentions about language.[12] Perhaps the most celebrated of these is his contention that there cannot possibly exist such a thing as a *private language.* Human beings can devise private codes, he granted. But, Wittgenstein says, any such code must be a translation of an existing public language. It isn't even theoretically possible, Wittgenstein thought, for there to exist a language that wasn't created by several interacting individuals.

Once such a language is created, Wittgenstein grants, a single person can use it on his own—he can go off into the woods and keep records with it. But any case of a single person's using a language for his own private use is, in

Wittgenstein's view, parasitic on there being, or once having been, a multiplicity of people using that same language.

Wittgenstein isn't making the pedestrian point that, as a matter of psychological fact, it would be hard, maybe even impossible, for somebody to single-handedly create a language (other than one that was a translation of an existing language). He's saying that it's logically impossible—that such a language is a ***surd,*** like a square circle.

Wittgenstein also held that one can't think unless one knows a language.[13] Wittgenstein seemed to grant that one could have mental states without knowing a language. But he pretty clearly thought that knowledge of a language was a prerequisite for any sort of reasoning.

If Wittgenstein is right, one cannot think unless one is embedded in a society of some kind. It is logically, and not just psychologically, impossible, Wittgenstein thought, for a human being who isn't either a current or erstwhile member of a society to think. These contentions, along with Wittgenstein's arguments for them, are put forth in his book *The Philosophical Investigations.*

Wittgenstein provides an interesting argument for his view that there cannot be a private language, which is known, appropriately enough, as the Private Language Argument (PLA). We'll discuss PLA in a moment. Wittgenstein provides two distinct arguments for the position that one cannot think without knowing a language. Because of the similarities between them, contemporary authors tend not to distinguish these arguments from one another, and they are collectively referred as to the Rule Following Argument.

6.1 Wittgenstein's first Rule-Following Argument

> (RF_1[14]) Not all ideation is ratiocinative. For ideation to be ratiocinative, it must be rule-governed. That is, it must be guided by a knowledge of the canons of logic. Thus, thinking involves following rules.
>
> But there is no psychological condition or event that necessarily accompanies an act of following a rule. To see this, go ahead and

follow some rule. (Add two multi-digit numbers together; read some passage out loud.) When you introspect, you may find that you are experiencing these or those feelings or images. But you could obviously follow the rule in question without having such experiences. Thus, following a rule is not a psychological act.

Contrary to what this argument assumes, not all mental states are images or sensations.[15] Your belief that 1 + 1 = 2 isn't such a thing; and your deploying that belief on some specific occasion isn't identical with your experiencing some image or sensation or series of images or sensations. Not everything in your mind is a fleeting conscious event. Much of what constitutes your mind consists of stable, enduring structures. Your belief that snow is white isn't a flash in the pan. It's been there for a while, and will continue to be there for a while and though it gives rise to momentarily conscious events, it isn't identical with any such event or with any aggregate of such events. Contrary to what Wittgenstein alleges, given only that there is no tickle or itch or other fleeting mental event that necessarily accompanies an act of rule-following, it doesn't follow that rule following isn't a psychological act.

6.1.1 Wittgenstein's second Rule-Following Argument

(RF_2) Consider the sentence "rabid dogs foam at the mouth." For argument's sake, let's suppose that there exists some entity M that is the meaning of this sentence. Let's also suppose that, without grasping or operating with the sentence "rabid dogs foam at the mouth" or any other expression, you behold M—you *grasp* it. You *see* it—through your "mind's eye," or some such. Finally, let's suppose that you not only grasp M, but know it to be true. In other words, you know that rabid dogs foam at the mouth.

If you can't see that, given M, Fido can't be rabid without foaming at the mouth, your grasp of M is, to that extent, useless and as good as non-existent. Thus, so far as you don't grasp how M bears on other meanings—so far as you don't grasp its *significance,* in other words—your grasp of M is useless and might as well not exist. The problem we had with the sentence "rabid dogs foam at the mouth" arises in

connection with M. Wanting to explain what it is to understand that sentence, we posited M. But it turned out that M is just like that sentence: it's useless unless *its* meaning is grasped. M turned out to be just another symbol—just another cognitively dead symbolic intermediary.

Meanings are themselves symbols. So unless we already know what it is to understand symbols, positing symbol-meanings doesn't help us understand the nature of symbol-comprehension. Positing symbol meanings in the hopes of explaining symbol-comprehension is like positing little people inside people's heads in the hopes of explaining how people think.[16] Positing these little people is useless unless it's *already* understood how people think. Positing meanings is useless unless it's *already* known how symbols are understood.

This argument of Wittgenstein's bears a striking resemblance to Aristotle's Third Man Argument (TMA).[17] The purpose of TMA is to undermine Plato's contention that spatiotemporal entities are instances of properties and, in addition, that the latter are not spatiotemporal. It will help if we take a moment to consider Aristotle's argument (it will soon be clear why the third and fourth sentences are italicized):

(TMA) For argument's sake, suppose that there is such a thing as the property of being a triangle. Let T_1 be that thing. If Platonism is right, *T_1 must itself be a triangle. Further, if Platonism is right, any case of two objects' sharing some property is a case of there being some third object of which they're both instances.* Thus, if T_2 is some triangle, there must be some third object T_3 such that each of T_1 and T_2 is an instance of T_3. But, by the same logic, there must be some fourth thing T_4 such that each of T_1, T_2, and T_3 is an instance of T_4. And so on *ad infinitum.*

Aristotle's argument assumes that properties are instances of themselves —that the property of being a triangle is itself triangular. But that's precisely what Platonism denies.[18] Platonism is the doctrine that spatiotemporal things

are but *instances* of non-spatiotemporal things, the latter being properties. The property of being spatiotemporal is therefore itself non-spatiotemporal and, therefore isn't an instance of itself. And the property of being a cat obviously isn't a cat, since cats are spatiotemporal, whereas that property is not.[19]

Just as, according to Aristotle, Platonism is struck with the problem of explaining why the property of being a triangle was itself a triangle, so according to Wittgenstein, those who believe in meanings are stuck with the problem of explaining how meanings are understood.

But meanings are not themselves understood. They're no more understood than the property of being a cat is a cat. Consider the expression "I." I'm referring, not to specific occurrences (utterances, inscriptions) of that expression, but to that expression itself—the ***expression-type,*** in other words, as opposed to its ***tokens.*** The meaning of that expression is some rule that assigns referents to its tokens depending on the circumstances. More precisely, the meaning of "I" is a rule to the effect that a token *t* of that expression that is uttered by x refers to x. My knowing that "I" has that meaning is part of what enables me to understand utterances of "I'm getting really bored," "I really like snow-boarding," etc. But that meaning is not itself understood. Meanings are either grasped or they're not. There's no such thing as grasping a meaning but not understanding it.

A consequence is that meanings don't have to be *interpreted.* It's incoherent to suppose that Hoigaard (who, let's assume, is Norwegian) should be able to grasp this meaning, but fail to see that people were referring to themselves when they said "I." If, upon hearing you say, "I am cold," Hoigaard has to thumb through his English-to-Norwegian dictionary to figure out what "I" means, then he doesn't know the meaning of the word "I."

This isn't to say, what would clearly be false, that one automatically recognizes every consequence of every principle that one grasps. People fail to see the consequences of the principles they grasp. In fact, since any given proposition has infinitely many consequences, people fail to grasp infinitely many consequences of anything that they grasp. But *to the extent* that one's grasp of a given principle is determinative of what one knows, there isn't a gap between grasping meaning, on the one hand, and applying it, on the

other. And to the extent that one doesn't see the consequences of some meaning or principle that one grasps, it isn't because one has failed to *interpret it.*

A story will clarify these points. Brown is a highly (but not superhumanly) intelligent person who not only speaks perfect English, but is exceptionally (but, again, not superhumanly) adept at understanding highly intricate sentences. XYZ is a sentence of English that is 100,000 words long but is otherwise normal. XYZ is grammatical, coherent, etc.[20] Even though Brown knows the relevant semantic rules, he obviously can't understand XYZ. But the problem isn't that Brown can't interpret the relevant semantic principles. The problem is that, because that sentence is so long, he can't *deploy* his semantic knowledge. In and of itself, that semantic knowledge is good to go. But, under the circumstances, it's weighed down by limitations on Brown's part that have nothing at all to do with his *semantic* competence. Brown's memory isn't perfect; he can't keep too many tasks going in his mind simultaneously, even if any given one of them is elementary. These aren't *semantic* limitations on Brown's part. The problem isn't that he doesn't know the relevant semantic rules. Nor is it that he hasn't adequately *interpreted* the semantic rules that he grasps. The problem is that his grasp of those principles is being inhibited by shortcomings that in no way detract from his semantic competence, even though they *do* limit his ability to *deploy* it. *To the extent that* what Brown knows is determined by his grasping those semantic rules, Brown *does* understand the expressions to which they assign meaning. But, of course, what Brown knows is a function of many things besides his grasp of those principles.

Wittgenstein's idea that meanings, if grasped, must be *interpreted* reflects his erroneous belief that, whenever one can't see a consequence of some principle that one grasps, it's because one hasn't adequately interpreted that principle. But principles aren't interpreted. They're grasped. And, if grasped, they guide judgments. And to the extent that one's grasp of a principle is determinative of what one knows, those judgments are consistent with those principles. But, of course, one's grasp of a given principle is but one factor among many determining the course of one's thought-process; so one's grasp of that principle may be prevented from eventuating in the judgments in which it would otherwise eventuate. But when that happens, it

isn't because that principle wasn't adequately *interpreted.*

6.2 The Private Language Argument

Neither of Wittgenstein's attempts to show that thinking isn't a psychological act is probative. They're interesting arguments, if I've interpreted them correctly; and if I haven't, it's an interesting question how they ought to be interpreted. But given what we've seen, we have no reason to reject the plausible, if not self-evident, position that thinking is a psychological act.

But, of course, Wittgenstein doesn't see it this way. Believing himself to have definitively established (the radically absurd principle) that rule-following and, more generally, thought aren't mental processes, he attempts to show that there can be no private languages. Here is his argument:

> (WA[21]) Imagine the following. Smith is alone on a desert island. He's always been this way and, therefore, has never learned a language. But he wishes to create one. He thus decides to stipulate that the meanings of expressions $X_1...X_n$ are to be respectively, *tree, berry, sand, sun,* etc. (He also specifies how those symbols are to be written down. For some reason, he has a notebook and a pen.) He does this all on Monday. We'll refer to the language Smith creates as "Smithese$_1$." One of the semantic rules for Smithese$_1$ is to the effect that:
>
> X_{35}: means that the berries on the dark side of the hill are poisonous.
>
> That same day, Smith sees some poisonous berries on the dark side of the hill, and writes X_{35} down in his notebook. (Let "X_{35}" be our symbol for that particular inscription.) On Tuesday, he reads what he wrote down. Question: is there any way that Smith can misinterpret what he's written? If he takes "X_{35}" to mean that the sun is out, his interpretation of "X_{35}" will be inconsistent with his Monday intentions. But so what? Since Smith is the only person who speaks Smithese$_1$. So the expressions belonging to it mean whatever he thinks they mean. An English-speaker can't just decide that "snow is white" means that grass is green, since it is

the thoughts (and practices) of people in general, and not of this or that specific person, that decides what is meant by "snow is white." But in this context, Smith *is* "people in general." So his believing that "X_{35}" means that the sun is out is like *everybody's* deciding that "snow is white" means that grass is green. And just as "snow is white" *would* mean that if everybody thought it did, so X_{35} means that the sun is out because, in effect, everybody thinks it does, the reason being that, in this context, Smith *is* everybody.

Thus, the expressions of Smithese$_1$ mean whatever Smith thinks they mean. This means that he can't be wrong as to what they mean.

But a symbol that can't be used wrongly doesn't mean anything. What makes utterances of "John is a soccer player" meaningful is that they are used wrongly if John isn't a soccer player. If "John is a soccer player" couldn't be used wrongly, you couldn't conclude anything from the fact that such an utterance was correct—you would be no less entitled to conclude that the moon was made of styrofoam than you would to conclude that the dog ate your wig.

Thus, there are languages only where there are right and wrong ways to use expressions. Since it is only where several people are using a language that there are right and wrong ways of using its expressions, it is only where there are several people that there are languages: all languages are public.

WA is spurious and establishes nothing. The reason that expressions of English (or any other public language) are meaningful is that people remember what they mean. If everybody forgot what "snow" meant, it wouldn't be of much use. To be sure, where English is concerned, no one person's forgetting the meaning of a word is going to strip of it meaningfulness, since there are millions of others who still know it. But that only means that English is causally more deeply entrenched than a language, like Smithese$_1$, that only one person speaks. This is not to mention that, if *I* forget the semantic rules for English, it will be as useless to me as Smithese$_1$ would be to Smith if he forgets the semantic rules of Smithese$_1$. By the same token, if Smith remembers the semantic rules for Smithese$_1$, the expressions

of that language *will* be useful for him. He'll know not to eat the berries on the side of the hill; he'll know where the good-tasting aardvarks are; etc.[22]

What about Wittgenstein's point that the expressions of Smithese_1 mean only what Smith thinks they mean? A language *is* a set of semantic rules. If, on Monday, "X_{35}" means that the berries on the dark side of the hill are poisonous, whereas on Tuesday that same inscription means that the sun is out, then the language Smith is using on Tuesday is *ipso facto* a distinct one from the language he is using on Monday. On Tuesday, he's speaking Smithese_2, not Smithese_1. There, "X_{35}" means that the sun is out *only* if it isn't an expression of Smithese_1. The expressions of Smithese_1, then, don't mean whatever Smith thinks they mean. If his opinions assign new meanings to $X_1...X_n$, those opinions are creating a new language that happens to be orthographically and phonetically coincident with Smithese_1; those opinions aren't assigning new meanings to expressions constituting some existing language. So Wittgenstein is just wrong to say that the expressions of Smithese_1 mean whatever Smith thinks they do.[23]

Thinking seems to involve following rules. Wittgenstein grants that people think. But since he believes totally absurdly, that rule-following and, therefore, thinking aren't psychological acts, he is forced to come up with a different analysis of what thought is. He wrongly says that thought consists in the manipulations of symbols. Given his belief that there aren't private languages, he holds that to think is to manipulate symbols of public languages.

Given that RF and WA are spurious arguments for false conclusions, Wittgenstein has given us no good reason to accept this position. In any case, we know it to be an erroneous one. A prerequisite for speaking a language is knowing its semantic rules and being able to operate in an intelligent and, consequently, thought-mediated fashion.

Chapter 6

Russell's Improvements on Frege's Work

1.0 Why definite descriptions aren't singular terms

In Chapter 1, we saw how Frege's groundbreaking insights into language and logic made it possible for him to solve otherwise unsolvable philosophical puzzles.

Bertrand Russell (1872–1970) extended Frege's work. Russell also showed that some important aspects of Frege's semantics are deeply incoherent, and Russell figured out how to eliminate those incoherencies without in the process sacrificing Frege's groundbreaking insights.

Russell's famous *Theory of Descriptions* was one of his earliest and most important extensions of, and improvements on, Frege's work.[106] The theory of descriptions concerns ***definite descriptions.*** A definite description expression of the form ‹the phi›, where phi is in the singular. So "the inventor of Velcro" is a definite description, as is "the inventor of bifocals." But "the people over there" is not a definite description, since "people" is a plural noun.

Definite descriptions are in the same grammatical category as "Mary" and "Bob" and other proper names. (Note: in this book, very generic proper names—e.g., "Smith," "Mary," "Bob"—will often be used. Always assume, for argument's sake, that they are unambiguous.) Thus, in respect of its grammatical form:

(V): "The inventor of Velcro is intelligent"

is just like

(B) "Bob is intelligent."

But we've seen that grammatical form and logical form sometimes pull apart. To determine whether this is going on here, let's take a careful look at what V says.

V clearly isn't true if nobody invented Velcro. V isn't true if several

people invented it. (If nine people invented Velcro, one could be *an* inventor of Velcro, but *the* inventor of it.) Finally, any inventor of Velcro must be intelligent. So if somebody did single-handedly invent Velcro, V is false if that person is unintelligent and true if that person *is* intelligent.

So here's what V says: at least one person invented Velcro; at most one person did so; and any such person is intelligent. Thus, V says that:

(V*) somebody uniquely invented Velcro, and any such person is intelligent.

In general[107]:

(DD) ‹ the phi has psi›

says that

(DD*) something uniquely has phi, and any such thing is psi.

Even if Bob invented Velcro, (B) and (V) say different things. V would be *false* in a situation where Bob was intelligent and where some unintelligent person had invented Velcro. But B would be true in such a situation. B and V thus have different ***truth-conditions.*** In other words, the conditions that must hold if the one is to be true don't coincide with those that must hold if the other is to be true. They aren't true under the same conditions. Therefore, they don't mean the same thing; they have different ***semantics.***

An expression's ***semantics*** is its meaning. A ***semantic rule*** is one that assigns a meaning to an expression.

1.1 The differences between proper names and definite descriptions[108]

There is some individual x, namely Bob, such that to know how to use "Bob" in a sentence—such that, in other words, to know the semantic rule for "Bob"—one must know that "Bob" refers to x. Thus, there is some individual x such that ‹Bob has psi,› for any property psi, is true iff x has psi.

But there is *no* individual x such that ‹the inventor of Velcro has psi› is true iff x has psi. It doesn't matter what property psi is. But to fix our ideas, let psi be the property of being 7-feet tall, and suppose that Bob is the *actual* inventor of Velcro (that he's the person who, in our world, invented Velcro);

and suppose that Bob is 7-feet tall. In that case, ‹the inventor of Velcro has psi› is true.

In light of this, let W be a world that is semantically just like ours—a world where people use the same expressions that we use and use them in accordance with the same rules. Suppose that, in W, Bob is the inventor of Velcro, and Bob is under 7-feet tall. In that case, ‹the inventor of Velcro has psi› is false, even though that sentence has the very same semantics that it has in our world. Thus, there is no individual x such that, to know the semantics of "the inventor of Velcro" is to know that it refers to x. Obvious generalizations of this argument show that, for any property phi, there is no individual x such that the semantic rule for ‹the phi› is to the effect that it refers to x.

1.2 Definite descriptions not referring terms

Definite descriptions aren't referring terms. No definite description refers to anything. The meaning of "the inventor of Velcro" isn't given by saying who it refers to, but by saying what whole sentences it mean. What whole sentences containing it mean is given in at least some cases by the rule: "the inventor of Velcro has psi" says that something uniquely invented Velcro, and any such person has psi." (Why "in at least some cases" as opposed to "always''? Because, when it occurs in some contexts, which we'll consider shortly, the relevant contextual definition isn't the one just given.)

In general, for any property phi, ‹the phi› is defined, not denotatively, but contextually—that is, by saying what whole sentences containing it mean. And what whole sentences containing it mean is (in general, though not, as we'll see, in all cases) given by the rule: ‹the phi has psi› is true if there exists something uniquely having phi, and any such thing has psi.

1.3 Contextual definition and the semantics of definite descriptions

Where definite descriptions are concerned, grammatical and logical form pull apart. "Jane is tall" (JT) has the same grammar as "the captain of the volleyball team is tall" (CT). But the propositions they express are structurally very different. JT is, whereas CT is not, such that, for some person x, it is true just in case x is tall.

Like the semantic rule for "someone," the meaning of "the inventor of Velcro" is given by saying what whole sentences containing it mean; and in at least some cases ‹the inventor of Velcro has psi› is true if the property of being a Velcro-inventor is uniquely instantiated and, moreover, that property has one instance in common with the property of being a psi.

1.4 The Theory of Descriptions defined

This theory of Russell's is known as the "Theory of Descriptions" (TD). The essence of TD is given by two assertions: (i) definite descriptions don't refer to anything; and (ii) in virtue of containing a definite description, a sentence attributes properties to other properties, not to individuals.

"The inventor of Velcro" doesn't refer to anyone or anything. "The inventor of Velcro is seven-feet tall" doesn't attribute the property of being 7-feet tall to anyone. That sentence says that *the property of being a unique inventor of Velcro* is instantiated and, in addition, that *the property of being a 7-feet tall Velcro-inventor* is also instantiated. So "the inventor of Velcro is seven-feet tall" is about properties, not individuals. It attributes the property of being instantiated to each of the two properties just mentioned.

2.0 Some arguments for the theory of descriptions: belief reports

Given only what we've said about TD, it might not seem like a very important theory. But it opens up new vistas, not just in semantics, but in metaphysics, epistemology, and the philosophy of mind.

Consider the sentence:

(PFY) "Ponce de Leon believes that the fountain of youth exists."

PFY is true. (To avoid irrelevant subleties relating to the use of past-tense markers, we'll assume, falsely but innocuously, that Ponce de Leon is still with us.) It's a simple historical fact. But does "the fountain of youth" refer to anything? No. There is no fountain of youth. For argument's sake, suppose that "the fountain of youth" refers to something. It follows trivially that there exists some object such that "the fountain of youth" refers to that object. Given what is meant by "fountain," "youth," etc., that object must be a

fountain of youth—that is, it must be something whose waters, when inbibed, give one eternal youth. But there is no such thing. So "the fountain of youth" isn't a referring term.

But now we have a problem. PFY is true. And it's therefore meaningful. (Nothing can be true or false without being meaningful.) Before Russell, most philosophers didn't want to give up on the idea that "the fountain of youth" *does* refer to something. They saw that it was meaningful, and they assumed, given its grammatical role, that it had to refer to something. This left them with the question: What does "the fountain of youth" refer to? As you might expect, a lot of them said that it exists "as an idea."

This doesn't work. Fountains are not ideas. When people disagree about whether there is a fountain of youth, they aren't disagreeing about whether some *idea* exists. I believe that Ponce de Leon was wrong. I believe it because I think that there is no actual fountain having the needed characteristics. For all I know, Ponce de Leon and I might see eye to eye when it comes to what *ideas* exist.

Maybe when people say that the fountain of youth exists *as an idea,* they only mean that we have an idea of such a thing. To have an idea of such a thing is to know what such a thing *would* be like, *were* it to exist. Thus interpreted, "the fountain of youth exists *as an idea*" is just another way of saying: "there is no fountain of youth (but people know what something would have to be like in order to be one)" which leaves us no further along than before.

Alexius Meinong (1853–1920) proposed a different solution.[109] He said that, although it doesn't exist, the fountain of youth "subsists." For a thing to subsist is for it to have some status intermediate between existence and non-existence.

This doesn't work. First of all, something either exists or it doesn't; there is no intermediate condition. Second, even if there is, it doesn't help, since subsistence is, by Meinong's own hypothesis, a way of failing to exist, which leaves us no further along than before.

Russell straightened this whole mess out. Here's what he said. (What follows will involve a repetion of some points made in Section 1.0.) "The fountain of youth" is a definite description. Definite descriptions don't refer

to anything and neither, therefore, does "fountain of youth." Definite descriptions are to be defined contextually, and sentences containing them must be reparsed if their meaning is to be made clear.

If it's grammatical form were to be trusted,

(E_1) "the fountain of youth exists"

would attribute existence to some fountain. But, since it must be reparsed if its meaning is to be made clear, its grammatical form is not to be trusted. Appropriately reparsed, that sentence becomes:

(E_2) "the property of being a fountain of youth is uniquely instantiated."

The meaning of PFY is therefore:

(E_3) "Ponce de Leon believes that the property of being a fountain of youth is uniquely instantiated."

2.1 Some arguments for the theory of descriptions: negative existentials

Consider the statement:

(N_1) "the monster under little Timmy's bed doesn't exist."

N_1 is true. But if "the monster under little Timmy's bed" referred to something, it wouldn't be true, since there would then exist some monster under little Timmy's bed to which that definite description referred, in which case, of course, there would exist a monster under Timmy's bed.

What's going on? For now familiar reasons, N_1 must be reparsed if its meaning is to be exposed. The needed reparsing yields:

(N_2) "the property of being a monster under little Timmy's bed isn't instantiated."

That property exists. (All this means is that there is a way something *would* have to be in order to be such a monster.) But nothing has that property. N_2

doesn’t refer to anything that doesn’t exist. Problem solved.

This makes it clear how to deal with statements like “the monster under Timmy’s bed doesn’t exist.” The going line before Russell was to say that it says of some *object* that it doesn’t exist. But that makes no sense at all, since it’s the same as saying “there exists an object (viz. a certain monster) that doesn’t exist.” What it really says, Russell showed, is that a certain existent property (viz. that of being a certain sort of monster, etc.) has a very existent property (viz. that of being un-instantiated).

2.2 Some arguments for the theory of descriptions (continued): identity-statements

Let’s start with a few platitudes about reference. To say that expression R refers to object O is to say that, for any property phi, ‹R has phi› attributes phi to O. If you want to say of Benjamin Franklin (BF) that he’s bald, you can do so by saying “Benjamin Franklin is bald.” (Of course, BF is no longer with us. But to set aside some extremely intricate and, in this context, irrelevant issues relating to tense, let us pretend that Franklin is alive and that we can speak of him in the present tense.) And if you *couldn’t* attribute baldness to BF by uttering that sentence, then “Benjamin Franklin” wouldn’t refer to BF. If, in saying “Benjamin Franklin is bald,” one was attributing baldness to *Aristotle,* then, in that context at least, “Benjamin Franklin” would refer to Aristotle.

Let’s generalize these points. For any property phi, “Benjamin Franklin” refers to BF (and not to Aristotle, for example) because ‹Benjamin Franklin has phi› attributes phi to Benjamin Franklin; and if one *couldn’t* attribute phi to Benjamin Franklin by saying ‹Benjamin Franklin has phi› then “Benjamin Franklin” would not refer to BF. Further, if, in uttering that sentence, one was attributing baldness to Aristotle (but not BF), then “Benjamin Franklin” *would* refer to Aristotle (but not Benjamin Franklin).Thus, for “Benjamin Franklin” to refer to BF it is both *necessary* and *sufficient* that ‹Benjamin Franklin has phi› attribute phi to BF. The fact that “Benjamin Franklin” refers to BF is *identical* with the fact that ‹Benjamin Franklin has phi› attributes phi to Benjamin Franklin, for any property phi.

In general, expression E refers to O if and *only* if for any property phi ‹E

has phi› attributes phi to O. E's referring to O may be *identified* with E's being such that, in virtue of containing it, a sentence *ipso facto* attributes a property to O.

In light of these points, let us look at an age-old puzzle that Russell neatly solved. Benjamin Franklin (BF) invented bifocals and he was also the first postmaster general. Therefore:

(CB) the inventor of bifocals is identical with the first postmaster general.

It's also true, of course, that:

(TB) the inventor of bifocals is identical with the inventor of bifocals.

But, whereas TB is trivial, CB is non-trivial.

The non-triviality of CB is hard to explain if definite descriptions are supposed to be referring terms. As we just saw, "the inventors of bifocals" refers to BF exactly if "Benjamin Franklin has phi" attributes phi to BF. Given that "the inventors of bifocals" does in fact refer to BF, it follows that CB attributes the property of being identical with the first postmaster general to BF. And given that "the first postmaster general" refers to BF, it follows that CB attributes the property of being identical with the inventor of bifocals to BF. And given both of these facts (viz. the facts stated in the last two sentences), it follows that CB says of BF that it has the property of being identical with BF. Which is trivial.

Frege proposed a famous solution to this. But we're going to consider Russell's solution first, since it's easier to understand Frege's solution in terms of Russell's than *vice versa.*

According to the theory of descriptions (TD) ‹the inventor of bifocals has phi› says that something uniquely invented bifocals, and any such thing has phi. And, according to TD, ‹the first postmaster general has phi› says that something was a unique first postmaster general, and any such thing has phi. It follows that, if TD is right, CB says the same thing as:

(CB_3) "something uniquely invented bifocals, and any such thing is a unique first postmaster general; and something was a unique first postmaster general, and any such thing is a unique inventor of bifocals."

It is readily seen that CB_3 is more succinctly expressed as follows:

(CBR) “there is some one thing that uniquely invented bifocals and was also a unique first postmaster general.”

To sum up, if TD is right, then CB means the same thing as CBR. It is easily verified that, indeed, CB is true exactly if CBR is true. TD gets the right result. Given that CBR is non-trivial—given that historical knowledge is needed to know it—TD readily explains why CB is non-trivial.

2.3 Some arguments for the theory of descriptions: positive existentials

For argument’s sake, suppose there to be some non-existent object O such that

(E_1) “the fountain of youth exists”

says that O exists. In that case, there *ipso facto* exists nothing to which E_1 attributes existence. Thus, if it is to say anything at all, E_1 must attribute existence to some *existing* object. Thus, supposing that E_1 says of some object that it exists, a *pre*condition for E_1’s saying anything is that the object in question exist. But in that case, we could conclude from the mere fact that utterances of E_1 are *meaningful* that E_1 is true. Thus, a consequence of supposing that E_1 says of some object that exists is that E_1 is not only true, but trivially so.

But, in actuality, E_1 is both false and non-trivial. It’s false since there’s no fountain of youth. It’s nontrivial since, to know whether it’s true, is isn’t enough to know what it means—let alone to know that it has a meaning.

The right analysis is the one proposed by Russell. It says that the property of being a fountain of youth is instantiated. Though nothing has it, the *property* of being a fountain of youth exists. In general, ‹the phi exists› doesn’t make the absurd claim that some non-entity exists; nor does it make the trivial claim that some entity exists: it makes the non-absurd, non-trivial claim that the property of being a phi is instantiated.

3.0 The incoherencies in Frege's semantic system

Let us now consider how Frege deals with just-described puzzle concerning identity. This will expose some profound problems with Frege's semantics, which we will discuss.

Benjamin Franklin (BF) invented bifocals and he was also the first postmaster general. Therefore:

(CB) the inventors of bifocals is identical with the first postmaster general.

It's also true, of course, that:

(TB) the inventor of bifocals is identical with the inventor of bifocals.

But, whereas TB is trivial, CB is non-trivial.

A moment ago, we considered Russell's solution to this problem. Now we're going to consider Frege's solution. Frege's solution is much more congenial to commonsense. But, when scrutinized, it turns out *either* to coincide with Russell's solution *or* to be totally incoherent.

3.1 Sense vs. reference

Here is Frege's solution:

> (FR)[110] "The inventor of bifocals" is a referring term; it picks something out. That expression refers to a given thing if and only if that thing has the property of being a unique bifocal-inventor. If a given thing didn't invent bifocals then "the inventors of bifocals" doesn't refer to it. If lots of different individuals invented bifocals, then "the inventors of bifocals" doesn't refer to anything. (If ten people collectively wrote *War and Peace,* it would be wrong to speak of *the* author of that book.) And if a given thing *did* uniquely invent bifocals, then "the inventors of bifocals" *does* refer to it.
>
> So there are *two* dimensions to the semantics of "the inventors of bifocals." There is what it refers to: this is the referent of "the inventors of bifocals." (Sometimes "reference" is used instead of "referent.") And there's the property that a thing must have to *be* its referent; this property is the *sense* of that expression. The referent is BF, and the sense is the

property of being a unique bifocal inventor.

Analogues of these points hold with respect to "the first postmaster general." There is its referent (which is BF); and there is its *sense* (which is the property of being a unique first postmaster general).

The reason that TB is non-trivial is that, although the definite descriptions in it have the same referent, they have different senses. A comparison may help. Imagine that you're looking at a certain person in the distance. That person is in fact Ralph, but you don't know it, because he's so far away and also, let us suppose, because he is (for reasons that aren't important) wearing a suit of armor. You see that he is doing jumping jacks. Later Ralph is standing right in front of you and, of course, you see him. He's wearing a plaid shirt. Once again, he's doing jumping jacks.

The first visual perception has the same object as the second. And each of those perceptions says of that object he or she is doing jumping jacks. But those two perceptions pick out that object in different ways. In the first case, that object is presented to you as something that has one set of properties (it's far away, wearing a suit of armor, etc.); in the other case it's presented as having a different set of properties (it's nearby, is wearing a plaid shirt, etc.). And that's why, even though those perceptions have the same object, and even though both attribute the property of doing jumping jacks to that object, those perceptions tell you different things.

The relationship between "the inventors of bifocals" and "the first postmaster general" is to be understood along similar lines. They have the same object (referent); but they pick that thing out through different bodies of information. That's why ‹the inventors of bifocals has phi› and ‹the first postmaster general has phi› tell you different things, *even though* they attribute the same property (phi) to the same object (BF).[111]

3.2 The partial collapse of Fregeanism into Russellianism

This all seems not only plausible, but practically self-evident.

But it's false and, indeed, incoherent. Let's consider the two sense-

perceptions discussed in FR. The one perception tells you that there is something in the distance, wearing a suit of armor, etc., and that thing is doing jumping jacks. So the sense in which Ralph is the *object* of that perception is that your perception *describes* Ralph. Your perception says: "something or other is over there, in the distance, wearing a suit of armor, [etc.]." Who (or what) fits that description? In other words, who has the relevant properties? Ralph. So Ralph is the object of that perception in the sense it *describes* him.

The information borne by that perception has the form "something has such and such properties [it's wearing a suit of armor, etc.], and any such thing is doing jumping jacks." Thus, the content of that perception is given by an ***existence-claim*** (a claim of the form "something has phi"). Ralph uniquely satisfies that existence-claim. (In other words, he and he alone has the relevant property.) And it is in virtue of his doing so that he is the object of the perception in question.

The same thing *mutatis mutandis* is true of your second perception. The information borne by that perception is along the lines of "something has thus and such properties [it's standing nearby, is wearing a plaid shirt ,etc.], and any such thing is doing jumping jacks." Ralph is the object of that perception by virtue of the fact that he and he alone has the relevant properties—that he and he alone is a thing standing in the relevant place while wearing a plaid shirt, etc.

In general, when a given thing x is picked out *by way of* descriptive information, what's really going on is that some statement of the form "something has such and such properties" is being affirmed (not necessarily verbally), and x is what fits the description.

If we bear this point in mind, we find that, so far as it isn't totally incoherent, Frege's analysis collapses into Russell's. Consider the sentence:

(IB[112]) "the inventor of bifocals is tall."

By Frege's own admission, "the inventors of bifocals" picks something out if and only if something fits the relevant description (i.e., if and only if something is a unique inventor of bifocals). And if there is a unique inventor of bifocals, that thing, whatever it is, is the thing to which "the inventors of

bifocals" refers. Supposing

that there is such a thing and that x is it, IB describes x as being tall. It goes without saying that, if there is no unique inventor of bifocals, then there is no way that IB can be true. (Frege himself emphasizes this last point. And he's right. There's no way that "the elephant in the corner is eating peanuts" can be true if there's no elephant in the corner.)

The upshot is that IB is true if and only if:

(IB*) something uniquely invented befocals, and any such thing is tall.

But what does the Theory of Discriptions (TD) say about IB? According to TD, the meaning of IB is: *IB**. So Frege's theory collapses into Russell's.

By obvious extensions of this, it follows that, if Frege's theory is right, then CB's meaning is identical with that of CBR. So Frege's theory accounts for the non-triviality of CB in the same way as Russell's theory (as TD).

3.3 Why Fregeanism doesn't completely collapse into Russellianism

As we know, Russell's theory is pretty good. The same is therefore true of Frege's, at least to the extent that it coincides with Russell's. The problem is that Frege's theory doesn't entirely coincide with Russell's. Unlike Russell, Frege took it for granted that definite descriptions are just what, given their grammar, they appear to be: referring terms. We've just seen that, if Frege's own theory of why CB is non-trivial is to be correct, then TD must be the right account of how definite descriptions work. But the essence of TD is that definite descriptions are not referring terms. It would obviously be incoherent to accept TD while also believing that definite descriptions are referring terms. We've just seen that it is precisely this incoherence that Frege is accepting, however unwittingly, in holding that definite descriptions are sense-bearing referring terms.

3.4 The incoherencies in Frege's semantics

Frege's view that definite descriptions are referring terms that have both sense and reference is one that has a lot of intuitive appeal, and people will hold onto it unless given *very* strong arguments to the contrary. Since the

arguments just given may not be enough for some, here is another.

Consider the following four sentences:

(1) "Smith believes that *the inventor of bifocals is 6-feet tall.*"

(2) "Smith believes that *the inventor of bifocals is not 6-feet tall.*"

(3) "Smith believes that *the first postmaster general is 6-feet tall.*"

(4) "Smith believes that *the first postmaster general is not 6-feet tall.*"

Were Smith to believe both 1 and 4, he would not, at least not for that reason alone, be irrational. Of course, *given* the empirical fact that some one person uniquely invented bifocals and was a unique first postmaster general, one of those beliefs must be wrong. But taken by themselves, the italized parts of 1 and 4 are consistent with each other. At the same time, neither entails the other. In other words, each is compatible with the negation of the other. So supposing that Smith is rational and that 1 is true, it doesn't follow that 3 is true or that 3 is false.

For the time being, let's suppose with Frege that definite descriptions are referring terms. In that case, the definite description in 1 co-refers with the definite description in 4. They both refer to BF. 1 says of BF that he has the property of being believed by Smith to be 6-feet tall; and 4 says of BF that he has the property of being believed by Smith to *not* be 6-feet tall. So Frege's view that definite descriptions are referring terms has the false consequence that 1 and 4 cannot both be true if Smith is rational.

Here's another way of stating the problem. According to ***Leibniz's Law,*** if x is identical with y, then x has a given property if and only if y too has that property. Leibniz's Law isn't directly concerned with language.

But it has a semantic corollary, namely, that if two expressions co-refer, they can be inter-substituted *salva veritate.* This means that, if E and E* co-refer, and S is a sentence containing E, then replacing E with E* will result in a sentence that is true if S is true and false if S is false. (It's assumed that the expressions being inter-substituted belong to the same language, are inflected in the same way, etc.) "Cicero was a Roman orator" is true; and so, therefore, is "Tully was a Roman orator," given that "Tully" co-refers with "Cicero." "Cicero was a 19th-century logician" is false; and so, therefore, is "Tully was

a 19th-century logician." Another way of stating this principle is that "inter-substituting co-referring terms preserves truth-value." The original sentence ("there is water in my glass") is true (false) iff the new sentence ("there is H_2O in my glass") is true (false).

Supposing, as Frege does, that "the inventor of bifocals" and "the first postmaster general" are referring terms, then, given that those expressions co-refer, replacing the one with the other should preserve truth-value. But it doesn't. 1 may be false if 3 is true and *vice versa.*

TD solves this problem. According to TD, the meanings of 1–4 are, respectively:

(1_R) "Smith believes that *something uniquely invented bifocals, and any such thing is 6 ft tall.*"

(2_R) "Smith believes that *something uniquely invented bifocals, and any such thing is* ***not*** 6 ft tall."

(3_R) "Smith believes that *something was a unique first postmaster general and any such thing is 6 ft tall.*"

(4_R) "Smith believes that *something was a unique first postmaster general and any such thing is* ***not*** 6 ft tall."

1 doesn't entail 3, and 1 is compatible with 4. 2 doesn't entail 4 and 2 is compatible with 3. (This is the hard data.) 1_R doesn't entail 3_R, and 1_R is compatible with 4_R. 2_R doesn't entail 4_R and 2_R is compatible with 3_R. So TD is compatible with the hard-data, whereas Frege's theory is not. Frege was aware of the problem just described, and he added an epicycle to his theory that he hoped would save it. It did not, as we'll now see.

3.4.1 Frege's cover-up: the creation of intensional contexts

Frege made profound contributions to logic and the philosophy of language. None of these disciplines even existed in anything like their modern form before he invented them, and most of the principles that he laid down in doing so hold up.

But Frege, like everyone else, had his less than stellar moments. Unfortunately, his blunders are treated with the same reverence as his discoveries. For this reason some very wrong and incoherent ideas of his are alive and well and continue to retard intellectual progress.

One of the worst of these is Frege's distinction between "extensional" and "intensional" contexts.[113] An extensional context is an expression that is "substitution tolerant." An expression E is substitution-tolerant if replacing any one of the referring terms occurring in it with a co-referring term yields a true sentence iff the sentence that hosted E was true and a false sentence iff the sentence hosted E was false. So "Smith is tall" is such a context, since replacing "Smith" a co-referring term will result in a sentence that has the same truth-value as the first. An intensional context is one which is substitution-*in*tolerant.

A story will help us move forward. The prettiest woman on the planet is also the most evil person on the planet. Smith wants to marry the prettiest woman on the planet, whoever that might be, but wants nothing to do with anyone who is evil. In that case:

(1) "Smith wants to marry the prettiest woman on the planet"

is true, but

(2) "Smith wants to marry the most evil person on the planet"

is false. And, of course,

(3) "it is necessarily the case that the prettiest woman on the planet is identical with the prettiest person on the planet"

is true, whereas

(4) "it is necessarily the case that the prettiest woman on the planet is identical with the most evil person on the planet"

is false.

According to Frege, 2 is what results when a referring term in 1 is replaced with a co-referring term, and 4 is what results when a referring term in 3 is replaced with a co-referring term.

But this isn't what is going on. What (1) really says is that:

(1*) Smith wants it to be the case *that Smith is married to a woman who is*

prettier than any other.

And what (2) really says is that

(2*) Smith wants it to be the case *that Smith is married to somebody who is more evil than anyone else.*

(1*) and (2*) thus make it clear what (1) and (2), respectively, are actually saying. In other words, (1*) and (2*) expose the *logical forms* of (1) and (2). Notice that, in (1), the grammatical object is an expression referring to a *person* ("the prettiest woman on Earth"), whereas in (1*) the grammatical object (the italicized part) is an expression referring to a *proposition.* And in (2), the grammatical object is an expression referring to a *person* ("the most evil person on the planet"), whereas in (2*) the grammatical object (the italicized part) is an expression referring to a *proposition.*

Thus, (1*) makes it clear that, if (1) is true, the object of Smith's desire is not a *person,* but rather a proposition. In other words, Smith wants a certain proposition to hold, namely, *that Smith is married to a woman who is prettier than anyone else.* Similarly, (2*) makes it clear that, if (2) is true, the object of Smith's desire is not a *person,* but a proposition, namely, *that Smith is married to somebody who is more evil than anyone else.* Since those two propositions *are* different, the fact that one of them has the property of being desired to be true by Smith, whereas the other does not lack that property, does *not* constitute a violation of Leibniz's Law.

Similar points hold in connection with 3 and 4. The real meanings of those sentences are, respectively:

(3*) it is necessarily the case *that any woman who is prettier than every other woman is prettier than every other woman.*

and

(4*) it is necessarily the case *that any woman who is prettier than every other woman is also more evil than every other person.*

(3*) attributes the property of being necessarily true to one proposition (the one meant by the italicized part), and (4*) expresses that same property to some *other* proposition. Since those propositions are distinct, the fact that

the one proposition is necessarily true doesn't entail that the same is true of the other; and there is thus violation of Leibniz's Law.

Thus, there are no intensional contexts. (In other words, all contexts are substitution-tolerant.) Whenever two expressions that appear to refer to the same *individual* cannot be inter-substituted, they actually refer to different *propositions.* Nonetheless, the term "intensional" is useful as a way of abbreviating otherwise hard-to- state truths. Thus used, an "intensional context" in which expressions that otherwise refer to individuals (e.g., "Benjamin Franklin") refer to propositions. An expression occurring in an intensional context does so ***intensionally.*** A context that isn't intensional is ***extensional,*** and expressions occurring in it do so ***extensionally.***

A brief review will help us identify one last problem with Frege's system. Frege's distinction between sense and reference embodies a failure to distinguish between reference and quantification. What Frege believes to be sense-bearing referring terms (e.g., "the person who invented Velcro") are quantifiers. (That is why "the man who invented Velcro is wealthy" is, by Frege's own admission, equivalent with "somebody or other uniquely invented Velcro, and any such person is wealthy"). Frege's distinction between intensional and extensional contexts embodies a failure to see that expressions that, given their grammatical properties, appear to refer to some one individual *in fact* refer to distinct *propositions.* Even though the inventor of bifocals is identical with the first post-master general,

(A) "Fred believes *that the inventor of bifocals is smart"*

may be true, even if

(B) "Fred believes *that the first post-master general is smart,"*

is false.

The italicized expression in A refers to one proposition; its counterpart in B refers to a different proposition. (We'll refer to the italicized expressions in A and B, respectively, as A* and B*). But B* is what results when a referring term is replaced with a co-referring term. According to the ***principle of special compositionality*** (PSC), A* and B* should co-refer.

PSC says that what an expression refers to depends on what its parts refer to. "Mary's favorite person" refers to Smith, whereas "Billy's favorite

person" refers to Jones. Why the difference in reference? Because "Mary" and "Billy" don't co-refer, and what a given expression refers to affects what expressions containing it refer to. For the same reason *mutatis mutandis,* if "Crusher" is Mary's nickname, and thus co-refers with "Mary," "Crusher's favorite person" co-refers with "Mary's favorite person." PSC is the obviously correct semantic corollary of the obviously correct principle that replacing equals with equals results in equals.

But Frege's semantics violates PSC. According to it, B* is what results when a referring term in A* is replaced with a co-referring term, and yet A* and B* don't co-refer. Russell's system doesn't have this problem. Russell said, correctly, that A* and B* are contractions of, respectively:

(A#) "some smart person single-handedly invented bifocals"

and

(B#) "some smart person post-master general before anyone else."

It didn't even occur to Frege to take anything at all like Russell's position. To compatibilize his system with PSC, Frege said that, when expressions occur in intensional contexts, they undergo reference-shifts. According to Frege, "the inventor of bifocals" *ordinarily* refers to Benjamin Franklin (BF); in other words, its so refers when it occurs in extensional contexts. But when it falls within the scope of one intensionality-inducing operator, it refers to the *sense* of "the inventor of bifocals." (An intensionality-inducing is what most other authos refer to as a "non-extensional connective." It is an operator having the property that any sentences or open-sentences falling within its scope are substitution-intolerant.) "Fred believes" is such an operator. Given that he insists on seeing definite descriptions as sense-bearing referring terms, Frege has no choice but to take this desperate measure: if he doesn't, his system violates PSC, in which case it's false, since PSC is true.

Thus, if Frege's system is right, then the occurrence of "the inventor of bifocals" in

(C) "Fred believes that Larry believes that the inventor of bifocals is smart"

refers, not to BF, nor yet to the sense of "the inventor of bifocals," but to the sense of an expression that denotes the sense of that expression.

For the same reason *mutatis mutandis,* a consequence of Frege's system is that no two of the occurrences of "snow" in

(D) "snow is white, but Brad thinks that snow is green, even though Brad is convinced that Mary thinks that snow is purple"

co-refer, and only one of those occurrences (the first one) refers to snow. In addition to being radically counterintuitive, Frege's position is logically and methodologically indefensible (or so I argue in Kuczynski (2007) and other places).

4.0 The theory of descriptions revisited: the wide-scope/narrow-scope distinction[10]

Let us refer to sentences containing definite descriptions as "D-sentences." So a D-sentence is any sentence of the form ‹...the phi...› There is a certain sub-class of D-sentences that are systematically ambiguous. One of the virtues of the theory of descriptions is that it, unlike its Fregean competitor, accounts for the fact that these sentences are ambiguous and it also accounts for the exact way in which they are ambiguous.

In this context, it will be helpful to remember that an ***open sentence*** is an expression that contains a free variable, and is therefore neither true nor false, but is otherwise just like a sentence. So ‹x is a human being› is an open-sentence; so is ‹x is even. › These open sentences are "true for" some values of their free variables. (The first is true for "Socrates," "Plato," etc. The second is true for "2," "4," etc.) But it must be emphasized that an open sentence is not itself true or false. (In this context, the word "sentence" will be short for "sentences or open sentences.")

An expression (or class of expressions[11]) is said to ***satisfy*** an open sentence if a true sentence results when that expression replaces the free variables in that sentence. Thus, "two" satisfies the open sentence "x is an even number," whereas "Mr. T" does not satisfy it, and neither does "three."

4.1 Scope, operators, and syntactic ambiguity

The semantics of these sentences is to be understood in terms of the concept of ***scope.***

Scope is a property of ***operators***—that is, of expressions that, given sentences or open-sentences, form new sentences. So "John believes that" is an operator, since, when joined with "grass is green," it yields the sentence "John believes that grass is green."

All ***quantifiers*** are operators. Examples of quantifiers are "all humans," "some triangles," "most dolphins." So are their counterparts in formalized extensions of English (e.g., "given any human being x," "for some triangle y," etc.). Why is "all humans" a quantifier? First of all, given a predicate (e.g., "are tall"), "all humans" yields a sentence ("all humans are tall"). And predicates are all open sentences (e.g., "are tall" can be thought of as synonymous with "x is tall"). Given that "all humans" can be thought of as synonymous with "given any human x," it's clear why "all humans" is an operator. When coupled with the predicate (= open sentence) "are tall" (= "x is tall"), "all humans" (= "given any human x"), yields the sentence "all humans are tall" (= "given any human x, x is tall").

All conjunctions (e.g., "and," "because," "or," "not") are operators. And, "because," and "or" are ***two-place operators,*** meaning that, given a *pair* of sentences as input, each yields a new sentence as output. Thus, given the pair of sentences ("grass is green," "snow is white"), "and" yields the sentence "grass is green and snow is white." Where "because" is concerned the pair must be ordered, since "grass is green because snow is white" doesn't have the same meaning as "snow is white because grass is green." It's a matter of debate whether "are" and "or" operate on ordered, or unordered, sentence-pairs. "Not" is a one-place operator. Given the sentence "snow is white," "not" yields a new sentence ("snow is not white").

In light of these points, consider the sentence:

(1) "In New York, someone gets mugged every five minutes."

This is ambiguous.[12] It could mean either:

(2) In New York, *given any given five-minute interval x,* **there is some individual y,** such that y gets mugged during x.

or

(3) In New York, **there is some individual y** such that, *given any given five-minute interval x,* y gets mugged during x.

(2) says that, in New York, no five-minute interval is free of muggings. It doesn't say that there is some *one* unfortunate individual who gets mugged every five minutes. But that *is* what (3) says.

The difference results from the fact that the relevant operators—the boldfaced and underlined expressions—are ordered differently. In (2), the operator "given any five-minute interval x" has ***wide-scope*** with respect to the operator "there is some individual y" (or, equivalently, the latter operator has ***narrow scope*** with respect to the former), whereas in (3) it's the other way around.

(1) isn't ambiguous for the same reason as "Jim is dumb" or "Sally is going to bank." Those sentences are ambiguous because the words "dumb" and "bank" are ambiguous. By contrast, none of the expressions occurring in (1) is ambiguous. (1) is *syntactically* ambiguous. It has multiple meanings because it stands for two distinct syntactic arrangements of expressions that are themselves unambiguous.

The stock example of syntactic ambiguity is:

(4) "everyone loves someone,"

which could mean either

(5) given any person x, there is some person y, such that x loves y,

or

(6) there is some person y, such that given any person x, x loves y.

(5) says that no person is without love for some person or other. It doesn't say that some *one* person is the object of every person's love. But that is what (6) says.

Grammatical surface structure makes it unclear how much scope operators actually have. For example, the scope of "because" in "snow is white because grass is green" includes *both* "snow is white" *and* "grass is green," even though grammatical surface structure would suggest otherwise. And the scope of "not" in "snow is not white" includes "snow is white." ("Not" = "it is not the case that." "Snow is not white" = "it is not the case that snow is white.") There would be no syntactically ambiguous sentences if grammatical surface-structure didn't fall short in this respect.

The fact that D-sentences are syntactically ambiguous can be understood in terms of these facts. The sentence:

(i) "Smith believes that the guy who is trying to kill him, after torturing him for several days, is well-disposed towards him."

This sentence is ambiguous. It could mean either:

(ii) Somebody x wants to kill Smith, after torturing him for several hours, and Smith believes that x is well disposed towards him

or

(iii) Smith believes that somebody x wants to kill him, after torturing him for several hours, and also that x is well disposed towards him.

(iii) has little chance of being true. People don't believe that people who are trying to torture and kill them are well disposed towards them. But it's easy to think of a situation where (ii) is true. Jerry is Smith's best friend. Jerry is actually very evil. But, like many evil people, he has mastered the art of appearing well-intentioned; and unbeknownst to Smith, Jerry wants to torture and kill him.

Whenever a definite description falls within the scope of a ***modal*** or ***epistemic*** operator, the result is a syntactically ambiguous sentence. Before Russell came along, such syntactic ambiguities, not being seen for what they were, led people to accept the most rank absurdities. Consider the sentence:

(iv) Jerry is looking for the fountain of youth.

Obviously there doesn't have to exist a fountain of youth for (iv) to be true. People look for things that don't exist. Russell saw this, and saw that, for this reason, (iv)'s meaning is:

(v) Jerry wants it to be the case that there exist a (unique) fountain of youth x, and, moreover, that he (uniquely) discover any such fountain.[13]

Pre-Russellian philosophers and contemporary non-Russellianism philosophers take (iv) to say that there exists a fountain of youth that Jerry is looking for. This is Nathan Salmon's (2005) position. Salmon qualifies his position by saying that the fountain of youth, though existent, is a "mythical" object. But, since mythical objects don't gush water, let alone immortality-giving water, Jerry won't consider his search to have to an end when he

discoveries the non-entity that Salmon describes. Given Russell's insight, there's no need to accept Salmon's rather doubtful position.

Because of his failure to grasp the implications of the wide-scope/narrow-scope distinction, another well known 20th century philosopher, W.V.O Quine (1908–2000), advocated the highly implausible view that "necessary," "possible," and modal terms in general, have no coherent meanings and that statements such as "it's possible that there will be rain" and "it's a necessary truth that triangles have three sides" are either false or meaningless.[14]

Here is his argument.[15] For argument's sake, suppose that the concepts expressed by "necessary" and "possible" are meaningful. In that case:

(FB) "it's a necessary truth that the inventor of bifocals was the inventor of bifocals"

expresses a truth, even though

(BF) "it's a necessary truth that the inventor of bifocals was the first postmaster general"

expresses a falsehood. Given the truth of FB, says Quine, it follows that the inventor of bifocals necessarily has the property of being identical with the inventor of bifocals. It follows by ***Leibniz's Law***[16] that the first postmaster general also has the property. But if he did, then BF would be true, which it is not.

With one exception, all of our assumptions were clearly correct. Leibniz's Law holds, and FB *is* a necessary truth if ever there was one. Since the only remaining assumption is that "necessary" and "possible" express coherent concepts, that assumption is wrong. It follows that all statements to the effect that something is possible or necessary or impossible (etc.) are either meaningless or false.

Arthur Smullyan (1948) saw that, given Russell's points about definite descriptions, there is no problem. BF is ambiguous between wide-scope and narrow-scope readings, these being:

(BFWS) it's a necessary truth that some one individual uniquely invented bifocals and, in addition, was the first postmaster general

and

(BFNW) some one individual x uniquely invented bifocals; and some individual y was uniquely a first postmaster general

and it's a necessary truth that $x = y$.

BFWS is straightforwardly false. BFNW is true. *As a whole,* it is not *necessarily* true. (It didn't *have* to be the case that some one person was a bifocal inventor and postmaster general.) But the underlined part *is* necessarily true: given any individual x, and given any individual y such that x is identical with y, it is a fact that x must be identical with y.[17]

Chapter 7

Logic, logicism, and the viability of attempts to mechanize thought

1.0 What did Frege do?

Pre-Fregean formal logic wasn't so much a discipline as it was a collection of rules of thumb. This was because Frege's predecessors weren't able to systematize the few results they were able to obtain, the reason being that they didn't see the principles underlying those results. But Frege saw those principles, organized those results, and added new ones of his own. In so doing, Frege released logic from the holding pattern it had been in for the previous two thousand years.

Unlike his predecessors, Frege saw that a sentence's logical form may diverge from its grammatical form. (In other words, he saw that a sentence S_1 may be such that what S_1 actually means is different from what S_1's grammatical structure suggests that it means.) And in many cases he figured out how to realign logical and grammatical form. (In other words, given many a sentence S_1 such that S_1's logical and grammatical forms diverge, Frege was able to produce a sentence S_2 such that S_2's logical form coincides

with S_1's and such that S_2's logical and grammatical forms coincide.) It was because of these insights of his that Frege, unlike all of his many predecessors, successfully ***formalized*** inferences involving ***quantified generalizations.***

But what does this mean exactly? What is it to "formalize" an inference, and what is a "quantified generalization"?

1.1 What is it to formalize an inference?

Any deductive inference can be thought of corresponding to a single sentence. For example, the inference from:

(a) "snow is white and grass is green"

to

(b) "grass is green"

corresponds to:

(c) "if snow is white and grass is green, then grass is green."

To formalize an inference is to show that the corresponding conditional is equivalent to a sentence that is an instance of an open-sentence all of whose instances are true. So in this particular case we need to find an open-sentence all of whose instances are true and of which (c) is an instance. Here is just such an open sentence:

(d) ‹if P and Q, then Q.›

(c) is equivalent with itself. (All sentences are self-equivalent.) (c) is an instance of (d). So we're done.

1.2 Frege on quantification[114]

The inference just formalized contained no sentences containing any quantifiers. Frege's successors had some success formalizing such inferences. They knew, for example, how to do what we just did. But they had little success with inferences containing sentences involving so much as a single quantifier (e.g., "all people smoke cigars"); and they had no success in the

way of formalizing inferences containing sentences containing multiple quantifiers (e.g., "one can fool all of the people some of the time, and some of the people all of the time, but not all of the people all of the time"). But Frege *did* succeed in formalizing such inferences.

We will henceforth refer to any inference involving at least one quantified sentence as a ***quantified inference.*** (So a quantified inference is an inference that involves so much as a single sentence containing a single quantifier.)

To move forward, we must note that Frege sees:

(SS) "someone snores"

as comprising the function ‹x snores› along with another function. This other function is a "second-order" function: it is a function that assigns objects to *functions.* Those objects are ***truth-values.*** (A sentence has the truth-value *true* if it's true and the truth-value *false* if it's false.) Frege sees ‹x snores› as expressing a function F that assigns truth or falsehood to a thing depending on whether that thing snores. And he sees "someone" as expressing a function G that assigns truth or falsehood to F depending on whether there is anything to which F, in its turn, assigns truth. If there is some x such that F assigns truth to x, then G assigns truth to F; otherwise G assigns falsehood to F.

This is a way of saying that "someone snores" is true just in case the property of snoring is instantiated. Since this is in fact precisely what SS says, Frege's analysis is correct.

Frege analyzes other quantifiers along similar lines. "Everything" expresses a function G that assigns truth or falsehood to a function F depending on whether there is something to which F assigns falsehood. If there is no object x such that F assigns falsehood to x, then G assigns truth to F; otherwise it assigns falsehood to F.

Consider the sentence:

(SI[115]) everything is self-identical.

Frege sees SI as saying:

(SI_2) **for any x,** *x is identical with x.*

The occurrences in SI_2 of "for any x" and "x is identical with x" correspond, respectively, to the occurrences in SI of "everything," and "is self-identical."

Frege sees the italicized part of SI_2 as expressing a function F that assigns truth to an object if that object is self-identical, and he sees the boldfaced part as expressing a function that assigns falsity to F if there is some individual to which F assigns falsity and that otherwise assigns truth to F. This is a way of saying that "everything is self-identical" says that the property of being self-identical is universally instantiated. Since this is, demonstrably, what that sentence is saying, Frege's analysis is correct.

A few final preliminary points: Frege chose to break up sentences containing the words "or" and "and." Thus, he saw "John is a tall mammal" as saying: John is tall and John is a mammal. And he saw: "John is male or female" as saying: either John is male or John is female. Finally, Frege chose to ***export*** the word sentence-internal occurrences of the word "not." Thus "snow is not white" becomes "it is not the case that snow is white." (In other words, he chose to rewrite such sentences as sentences in which (a) that word is replaced with "it is not the case that," and (b) "it is not the case that" is placed before the sentence being negated.)

By reparsing sentences along these lines, Frege was able to formalize an extensive class of quantified generalizations.[116] Some symbolic notation, some of it already familiar, will help us see how he did this. Let "→" be

defined as before. Let ‹P↔Q› be short for ‹P→Q and Q→P.› In other words, ‹P↔Q› means that P is a logical consequence of Q and *vice versa.* Let "~" mean "it is not the case that." Let "(x)" mean "given any object x." And, as before, we'll use parentheses to indicate how sentences are to be grouped together. So "Smith is tall and (grass is green or birds fly)" says that it's the case both that *grass is green or birds fly* and that *Smith is tall.* By contrast, "(Smith is tall and grass is green) or birds fly" makes the very different statement that *either* it's the case that *Smith is tall and grass is green or* it's the case that *birds fly.*

Bearing these points in mind, consider the following argument:

Argument #1

Premise: No mammal is intelligent if it has pointy ears.

Premise: Smith is a pointy-eared mammal.

Conclusion: Smith isn't intelligent.

This argument is valid. But, thus expressed, it isn't *formally* valid. This is because "no mammal is intelligent if it has pointy ears" has the same syntax as "Jones is intelligent if he has pointy ears," a consequence being that Argument #1 has the same syntactical form as the following, clearly invalid argument:

Argument #2

Premise #1: Jones is intelligent if he has pointy ears.

Premise #2: Smith is a pointy-eared mammal.

Conclusion: Smith isn't intelligent.

But by reparsing the sentences occurring in these arguments, Frege showed why it is that, their surface-structures notwithstanding, Argument #1 is valid whereas Argument #2 is not.

It will help if we make it clear at an intuitive level what the premises in Argument #1 are saying. The first premise is to the effect that anything that is an intelligent mammal does *not* have pointy ears. (In other words, given any object x, if x is an intelligent mammal, then x doesn't have pointy ears. To put it yet another way, given anything x, if x is a mammal, then, if x is intelligent, it follows that x doesn't have pointy ears.) The second premise is

to the effect that Smith is a mammal and Smith has pointy ears. Finally, the conclusion is to the effect that it's not the case that Smith is intelligent.

Thus, duly reparsed, Argument #1 becomes:

Argument $\#1_{RP}$[117]:

Premise #1: Given any object x, if x is a mammal, then, if x is intelligent, x doesn't have pointy ears.
Premise #2: Smith is a mammal and Smith has pointy ears.
Conclusion: It is not the case that Smith is intelligent.

Duly symbolized, Argument $\#1_{RP}$ becomes:

Argument $\#1_{S}$[118]:

Premise #1: (x)(x is a mammal→((x is intelligent)→~(x has pointy ears)).
Premise #2: Smith is a mammal and Smith has pointy ears.
Conclusion: ~(Smith is intelligent).

Replace the constants in Argument $\#1_{S}$ with variables. In other words, replace "mammal," "intelligent," and "has pointy ears" with (let us say) F, G, and H, respectively, and replace "Smith" with A. Make the needed grammatical adjustments. The result is:

Argument $\#1_{G}$[119]:

Premise #1: (x)(x has F→((x has G)→~(x has H)).
Premise #2: A has F and A has H.
Conclusion: ~(A has G).

Replace the variables in Argument $\#1_{G}$ with any (grammatically appropriate) constants you wish. The result will always be a valid argument. (It's assumed that the replacements are uniform—i.e., that you don't, for example, assign "intelligent" to one occurrence of "F" and "tall" to some other occurrence of it.) For example, replace F, G, and H, respectively, with "prime," "even," and "greater than 3"; and replace A with "7." The resulting

argument is:

Argument #1_{AA}[120]:

Premise #1: (x)(x is prime→((x is even)→~(x is greater than three)).
Premise #2: 7 is prime and 7 is greater than three.
Conclusion: ~(7 is even).

This argument is clearly valid. The only even prime is 2. So no even prime is greater than three. Given that 7 is prime and greater than three, it follows that 7 isn't even.

The result of uniformly replacing the variables in Argument #1_G with constants *always* yields a valid argument. Thus, Argument #1_G successfully formalizes Argument #1, the reason being that (so-called) Argument #1_G isn't really an argument at all, being instead an argument-*form,* all of whose *instances* are valid.

Now let's subject Argument #2 to the same treatment. Reparsed, it becomes:

Argument #2_{RP}

Premise #1: Jones is intelligent and Jones has pointy ears.
Premise #2: Smith is a mammal and Smith is pointy-eared.
Conclusion: It is not the case that Smith is intelligent.

To generalize argument #2RP, replace "mammal," "intelligent," and "has pointy ears" with F, G, and H, respectively, and replace "Smith" and "Jones" with A and B, respectively. Duly generalized, it becomes:

Argument #2_G:

Premise #1: B has G and B has H.
Premise #2: A has F and A has H.
Conclusion: It is not the case that A has G.

There are instances of 2_G that are invalid. Let A and B be the numbers ten and twenty, respectively. Let F, G, and H be the properties of being greater

than one, greater than two, and greater than three, respectively. The result is:

Argument #2_N:

Premise #1: the number twenty is greater than two and the number twenty is greater than three.

Premise #2: the number ten is greater than one and the number ten is greater than three.

Conclusion: It is not the case that the number ten is greater than two.

Since argument #2_N is spurious and has the same form as argument #2, the latter isn't formally (or otherwise) valid. Thus, Frege's reparsing of intuitively valid arguments, such as #1, and of intuitively invalid arguments, such as #2, enabled him to formalize those inferences. Once an inference is formalized, it can be carried out ***algorithmically*** and, therefore, without relying to any degree on *ad hoc* methods.

2.0 Logicism[121]

Frege's primary objective wasn't to formalize logic. It was to formalize arithmetical truth. Frege invented formal logic in order to do this. But what does it mean to say he wanted to "formalize arithmetic"?

The sentences by which arithmetical truths are ordinarily expressed are not formally true. For example, "2 + 2 = 4" has the same form as "2 + 2 = 5" and "1 + 1 = 3." Thus, "2 + 2 = 4" is not a formal truth. It's an informal, analytic truth, like "any case of knowledge is a case of true belief." Frege was keenly aware of this fact.

But Frege believed that, when their grammatical forms are brought into alignment with their logical forms, arithmetical truths turn out to be formal truths.[122] To establish this, he needed to identify a systematic way of translating informally true arithmetical sentences into formally true ones. So he needed to find a translation-rule R that, when given an informally true arithmetical sentence (e.g. "1 + 1 = 2"), paired it off with a formally true, but otherwise synonymous, arithmetical sentence.[123]

Thus, Frege believed that, as they are *conventionally* stated, arithmetical

statements, when true, are both informal and analytic. But he also believed that such statements are formal truths in disguise.

The position that arithmetical statements are formal truths (in disguise) is known as "logicism." Frege spent the better part of his career trying to prove logicism to be true.

In the process of trying to establish the truth of logicism, Frege made some important discoveries. One is of special importance. Statements that, given their surface-structures, *appear* to be about objects are really about properties. So "nobody snores" is about the property of snoring, and it attributes the property of being uninstantiated (of having a null-class extension) to that property.

This discovery of Frege's is closely related to, and almost coincident with, his discovery of the fact that grammar doesn't always reflect underlying logical structure.

2.1 Logicism (continued)

Analysis is elimination. To analyze causality is to show how statements of the form ‹x caused y to happen› can be translated into statements that don't contain the word "cause" or any other comparable expression (e.g., "force," "coerce"). Hume tried to do this. He argued that

(i) ‹x caused y to happen ›

is synonymous with

(ii) ‹x immediately precedes y and is adjacent with it, and the sequence consisting of those two events instantiates a general regularity. ›

Notice that (ii) doesn't contain "cause" or "force" or any other such term. (Unfortunately, this analysis is a complete failure. In Chapter 17, we will see why.)

Similarly, to give an *analysis* of the number one is to say how statements about it can be purged of any expression referring to it. Frege did this, and he did the same thing *mutatis mutandis* for statements about all other whole numbers.

On the basis of Frege's work, Russell showed how the same thing can be

done for fractions and irrationals.[124] We won't go into the specifics of how Frege and Russell do this, but we will discuss, in very general terms, some of the philosophically more important aspects of Frege's impressive achievement.

2.2 Frege's initial analysis of number-statements[125]

The statement:

(1) "there are two apples on the table"

is equivalent with:

(2) "*something x is an apple on the table;* **something y is an apple on the table;** *and x ≠y*; AND FOR ANYTHING Z THAT IS AN APPLE ON THE TABLE, EITHER Z = X OR Z = Y."

Let's look at (2) for a second. The first (italicized) part, guarantees that there is at least one apple on the table. The second (boldfaced) part does *not* guarantee that there is a second apple on the table, since it has not been ruled out that the apple in question is identical with the one described in the first part. But taken in conjunction with the third (italicized) part, the second part *does* guarantee that there are at least two apples on the table. But the first three parts don't guarantee that there are *only* two apples on the table; they guarantee only that there are *at least* two. But taken in conjunction with the fourth (capitalized) part, the first three parts *do* guarantee that there are exactly two apples on the table. Thus, (1) is equivalent with a sentence that contains no mention of the number two or, indeed, of any other number.[126]

By obvious extensions of this reasoning:

(3) there are three apples on the table

is equivalent with

(4) "something x is an apple on the table, and so is something y, and so is something z; and y ≠ x and y ≠ z and z ≠ x; and given anything w that is an apple on the table, w = x or w = y or w = z."

In Frege's (correct) view:

(5) "there are zero apples on the table"

is equivalent with

(6) "the property of being an apple on the table is uninstantiated,"

and

(7) "there is one apple on the table"

is equivalent with

(8) "something x is an apple on the table and, given anything y that is such an apple, y = x."

The expressions "two," "three," "zero," and "one" occur in sentences (1), (3), (5), and (7). Those sentences are equivalent with (2), (4), (6), and (8), respectively, even though no number-expression occurs in any of them. Thus, Frege has successfully *eliminated* any references to number occurring in (1), (3), etc., and therefore, has successfully analyzed them. But there's a problem, as we'll now see.

2.3 "Number" defined [127]

In (1), (3), etc. the expressions "one", "two", etc. are functioning as adjectives; and the method just described tells us how to eliminate those expressions from those sentences. In general, that method tells us how to eliminate number-expressions that are functioning as adjectives. But it doesn't tell us how to eliminate number expressions that are functioning as nouns. It doesn't tell us, for example, how to identify a sentence that is equivalent with "two is less than three" that doesn't contain "two" or "three" or any other number-expression.

Frege's brilliant solution to this problem is to be understood in terms of the fact that two sets have the same number of objects if they can be put into a one-one correspondence with each other.[128]

Let's suppose that Smith has five shirts and that Brown has five cars. Let S_1 be the set of Smith's shirts and let S_2 be the set of Brown's cars. S_1 can be

put into a one-one correspondence with S_2. In other words:

(1) there is a rule (or "function") that pairs off each member of each set with at least one member of the other set and pairs off no member of either set with more than one member of the other.

(1) says the same thing as:

(2) there is some function F such that, given any member x of either set, F assigns some member y of the other set to x and such that F assigns z to x iff z = y.

If two sets can be put into one-one correspondence, we'll say that they are "similar."[129]

Given any two sets, S_1 and S_2, the statement:

(a) ‹S_1 and S_2 have the same number of members›

is equivalent with:

(b) ‹S_1 and S_2 are similar. ›

Two sets are similar just in case they are both ***instances*** of the same number. A pair of shoes is an instance of the number two.

Anything of which there are instances is *ipso facto* a property. We may thus identify the number two with the property of being a pair of objects (i.e., with the property had by all and only pairs of objects).

A pair of objects is a set S such that, for some x and some y, x is a member of S and y is a member of S and $x \neq y$ and such that, for any z, if z is a member of S, then z = x or z = y.

Thus, the number two is identical with the property of being a set S such that, for some x and some y, $x \neq y$, and x is a member of S and y is a member of S, and for any z, if z is a member of S, then z = x or z = y.

Notice that this analysis isn't circular. "Two" is defined as "what all and only pairs of objects have in common," and our definition of "pair of objects" uses no number expressions.

Any given whole number is to be identified with a similar property. For

example, the number three is the property of being a collection of three things (i.e., it is the property had by all and only such collections). Though this definition seems circular, it isn't, since "collection of three things" can be defined without using the word "three" or any other number expression. To wit: S is a collection of three things iff there are objects x, y, and z such that x is a member of S and y is a member of S and z is a member of S; and such that $x \neq y$ and $y \neq z$ and $x \neq z$; and such that, for any w, if z is a member of S, then $w = x$ or $w = y$ or $w = z$. The number three is thus the property of being a set of the sort just defined.

The number one is the property of being a set S such that something x belongs to S and such that anything y that belongs to S is identical with x. The number zero is the property of being a set S such that nothing belongs to S.

In general, a whole number N is the property of being an N-tuple.

2.4 Numbers as properties of properties

In

(JG) "Jim has green apples,"

the word "green" picks out a property that can be had by individual apples. But in

(J_0) "Jim has zero apples,"

the word "zero" can't possibly pick out such a property. Let's suppose that J_0 is correct. In that case, Jim has *no* apples. *A fortiori* he has no apples having this or that property.

In J_0, the word "zero" denotes a second-order property—that is, it denotes a property of a property. More specifically, it denotes the property of being an uninstantiated property. Thus, J_0 says that:

(J_{0*}) The property of being an apple belonging to Jim has the property of being an uninstantiated property.

A moment's reflection makes it clear that J_{0*} and J_0 do indeed say the very

same thing.

Whenever number-expressions function as adjectives, they denote properties of properties. To take another example, in

(J_2) "Jim has two apples"

the word "two" couldn't possibly pick out a property that could be had by individual apples. Any *one* apple *ipso facto* isn't *two* apples. In J_2, the word "two" picks out a property of the property of being a property that has two instances. In other words, it picks out the property of being a pair. And J_2 says that

(J_{2*}) the property of being an apple owned by Jim has the property of being a pair.

Let's sum up. J_0 says that the property of being an apple owned by Jim is uninstantiated; J_1 says that it's uniquely instantiated; and J_2 says that it's instantiated exactly twice.

2.4.1 Numbers as properties of properties (continued)

The obvious thing to say about JG is that it attributes the property of being green to one or more apples belonging to Jim. But, as Frege pointed out, this isn't what it's doing. There is no specific apple x such that JG says that x is green. Given any apple x owned by Jim, JG could be true even if x was *not* green. This would happen if Jim had some *other* apple that *was* green. Thus, there is no apple x such that JG says that x is green. *A fortiori* JG doesn't say of each of several apples that it is green. Since there is no one apple owned by Jim to which JG attributes any property, JG isn't about apples owned by Jim. Rather, JG is about the *property* of being an apple owned by Jim; and JG says of this property that its extension overlaps with the extension of the property of being green. For this reason, JG's logical form is no more in alignment with its logical form than J_2' s (or J_0's) logical form is in alignment with *its* logical form.

2.5 Some logicist paraphrases of arithmetical statements

Frege's analysis of number (*a whole number n is the property of being an n-tuple*) makes it clear what is meant by statements such as:

(TLT) "two is one less than three (i.e., three is the successor of two)."

The number two is the property of being a pair. The number three is the property of being a triple. Thus, TLT says that:

(TLT*) given any instance C of the number three (i.e., given any triple), and given any member x of that triple, any class C* that doesn't contain x but is otherwise like C is an instance of the number two.

Although number expressions occur in TLT*, these are all easily eliminated. A "triple" is any class K such that there are objects x, y, and z, all distinct, belonging to K and such that anything w belonging to K is identical with x or y or z. Since, as we know, "the number three" is the property of being a triple, the occurrence of "the number three" in TLT* is just shorthand for a much longer expression that contains no number-expressions at all, the same being true, for similar reasons, of the occurrence of "the number two."

2.5.1 Adding whole numbers

Frege's analysis of number (*given any whole number n, n is the property of being an n-membered set*) makes it possible to translate arithmetical statements into statements in which no number-expressions occur.

"1 + 2 = 3" says that the union of a one-membered set and a non-overlapping two-membered set is a three-membered set. In general, ‹A + B = C› says that the union of an A-membered class and a non-overlapping B-membered class is a C-membered class.

Here's the idea. If Sally has one house, and Jerry has two yachts, then there are three things that are *either* houses belonging to Sally *or* yachts belonging to Jerry. If Larry has one pool and two cars, then there are three things that are either pools belonging to Larry or cars belonging to Larry. At the same time, if Larry has one expensive possession, and he also has two cars, one of which is expensive, there are only *two* things each of which is

either a car belonging to Larry or an expensive thing belonging to Larry.

This shows that, in general, if there's one phi and there are two psi's, then, *provided* that neither of the psi's is a phi, there are three things that are either phi's or psi's. Another way of putting this is to say that: given any one-membered class and any non-overlapping two-membered class, the class of things belonging either to the one class or the other is a three-membered class. And "1 + 2 = 3" is a compressed way of saying exactly this. "3 – 2 = 1" says the same thing, given that it is a notational variant of "1 + 2 = 3."

In general, each of ‹A + B = C› and ‹C – B = A› is a compressed way of saying that, for any properties phi and psi, given A phi's and B psi's, none of them a phi, there are C phi's or psi's.

The occurrences of "1," "2," and "3" in "1 + 2 = 3" are really adjectival. What is being said is that, given one *phi* along with two *psi's,* neither of which is a phi, there are three *phi's or psi's.*

Remember that, for any whole number N, it can be said what an N-membered set is without using any number expressions. Thus, despite first appearances, the just-stated analysis is not circular. The number two is the class of two-membered sets. A "two membered" set is one such that some member x of that set is not identical with some other member y and such that z belongs to that set just in case z is identical with x or y. No mention of the number two.

In general, for any whole numbers A, B, and C, ‹A + B = C› can be translated into statements containing no number expressions.

2.5.2 Multiplying whole numbers

"2 × 3 = 6" says that the Cartesian product of a two-membered set and a three-membered set is a six-membered set. (For reasons that will become apparent, it doesn't matter whether the sets overlap.) The Cartesian product of two sets S_1 and S_2 is the set of ordered pairs <a, b> such that a belongs to the one set and b belongs to the other.

Explanation: "2 × 3 = 6" can be thought of as saying that double-counting the members of a three-membered group yields a total count of six. In other words, given a group of three objects, if you count that group once and then count it once more, the total count will be six.

Let C_1, C_2, and C_3 be Sally's three cars, and let A_1 and A_2 be Larry's two apples. If A_1 is allowed to couple with each of Sally's cars, the result is the set containing the pairs $<A_1,C_1>$, $<A_1,C_2>$, and $<A_1,C_3>$. In this context, A_1 is applying a distinctive marker to each member of the smallest set containing each of Sally's cars. A_1 is marking C_1, C_2, and C_3 with the ordered pairs $<A_1, C_1>$, $<A_1, C_2>$, and $<A_1, C_3>$, respectively. When you count your socks, you apply a distinctive marker to each sock. You mark the first sock with a "one," the second with a "two," etc. Thus, A_1 is doing to each of Sally's houses what you are doing when you count your socks.

If Larry's other apple is allowed to do the same thing *mutatis mutandis,* the result is the set containing the following pairs:$<A_2,C_1>$, $<A_2,C_2>$, and $<A_2,C_3>$. Notice that none of the markers used by A_2 is identical with any of the markers used by A_1. Thus, the two counts will be independent, in the sense that if Z is a set containing the one count, and Z* is a set containing the other, Z won't overlap with Z*. Thus, union of Z and Z* can be thought of as the result of double counting Sally's houses. The members of that union-set are: $<A_1,C_1>$, $<A_1,C_2>$, $<A_1,C_3>$, $<A_2,C_1>$, $<A_2,C_2>$, and $<A_2,C_3>$. That union-set, which is the Cartesian product of Z and Z*, is a six-membered set.

It's obvious that, given any two-membered set and any three-membered set, the union of the headcounts collectively done by both members of the two-membered set will be a six-membered set. In other words, the result of double counting the three membered set will be a six-membered set. And this is just what "$2 \times 3 = 6$" says. "$6 \div 2 = 3$" says the same thing, given that it's a notational variant of "$2 \times 3 = 6$."

"$3^2 = 9$" says the same thing as "$3 \times 3 = 9$." Thus, "$3^2 = 9$" says that the Cartesian product of a three-membered set and a non-overlapping three-membered set is a nine-membered set.

In general, for any whole numbers A, B, and C, ‹$A \times B = C$› and ‹$A^2 = B$› can be translated into statements containing no number expressions.

It's obvious that statements about rational numbers (fractions) are equivalent with statements about whole numbers. In fact, they *are* such statements. To say that 1/2 is less than 3/4 is less than is simply to say that

one quotient is less than some other quotient. For any whole numbers A, B, and C, ‹A × B = C› is a notational variant of ‹A = C ÷ B›. Since, as we've seen, ‹A × B = C› can be translated into a statement containing no number-expressions, the same is true of ‹A = C ÷ B.›

2.5.3 Adding and multiplying reals

√2 is irrational. In other words, there are no whole numbers p and q such that √2 = p/q. But any given statement about √2, or any other irrational number, can be rewritten as a statement about rationals (and, therefore, can be rewritten as a statement about whole numbers and, therefore, can be rewritten as a statement not containing *any* number-expressions). Consider the statement:

(i) $a_1/a_2<\sqrt{2}$.

For (i) to be true, it is necessary and sufficient that a_1 and a_2 be such that $(a_1/a_2)^2<2$. Thus, the meaning of (i) is:

(ii) $(a_1/a_2)^2<2$.

By analogous reasoning, the meaning of

(iii) $a_3/a_4<\sqrt{3}$

is:

(iv) $(a_3/a_4)^2<3$.

It follows, given that no two reals are adjacent, that for

(v) $\sqrt{2}<\sqrt{3}$

to be true is for it to be the case that

(vi) there are whole numbers a_5 and a_6 such that $2<(a_5/a_6)2<3$.

Given any whole number n such that √n is irrational, and given *any* statement S of the form

(vii) ‹...√n...›,

there is an equivalent S* statement of the form:

(viii) ‹...n...›

Let us take stock. Statements about √2 can be represented as statements about the class all rational numbers whose squares are less than two. Thus, √2 may be identified with the class of all ratios R such that $R^2<2$.[130] Other irrationals are to be understood along similar lines. Thus, statements about irrational numbers can always be translated into statements about rational numbers.

Statements about rationals can obviously be translated into statements about whole numbers. Statements about whole numbers can be translated into statements that don't contain any number-expressions. (See Sections 2.5.1–2.5.2.) Thus, statements about irrational numbers can always be translated into statements that don't contain any number-expressions.

2.6 A problem with Frege's analysis

Though basically identical with the analysis given by Frege, the one given in Section 2.3 differs from Frege's in one important respect. Frege identified numbers, not with properties of sets, with *sets* of sets. Frege identified the number two with the set containing all and only those sets S such that, for some x and some y, x is a member of S and y is a member of S and x ≠ y and such that, for any z, if z is a member of S, then z = x or z = y. So, for Frege, the number two is the set of all sets of the sort just described. The number one is the set containing all and only those sets S such that something x belongs to S and such that anything y that belongs to S is identical with x. The number zero is the set containing all and only those sets to which nothing belongs.

The problem is that, if numbers are identified with sets rather than properties, it follows that the number two *could* have been something other than what it actually is. Consider some pair of objects (e.g., your two favorite socks) and let it be the set containing just those two things. Given a universe where those two socks don't exist, if K_2 is the class of all pairs in that

universe, then K_2 won't be identical with the class of all pairs in *our* universe. Since two sets cannot be identical unless they have the same members, it follows that *the number two could have something other than what it is,* which is absurd. The same thing *mutatis mutandis* holds of any other whole number: if a whole number N is identified with a set of sets, that whole number will be different things in different universes.

(The only exception to this is the number zero. If defined in the way that Frege proposes—that is, as the set of all sets S such that nothing belongs to S—it doesn't change from universe to universe. This is because, in order for two sets to differ, one of them must have a member not had by the other. This is known as the ***axiom of extensionality.*** Since no empty set has any members not had any other empty set, there cannot be two distinct empty sets. Thus, there is only one empty set; and this is true in every universe, not just ours).

But *what it is* to be a two-membered doesn't vary from universe to universe. In any given universe, for S to be a two-membered set is for it to be such that, for some x and some y, x is a member of S and y is a member of S and $x \neq y$ and such that, for any z, if z is a member of S, then $z = x$ or $z = y$. Thus, the *property* of

being a two-membered set doesn't vary from universe to universe. To that extent, it's better to identity the number two with the property of being a two-membered set than it is to identity with the set of all such sets.

But numbers can still be identified with sets of similar sets. The number two can be identified with sets of all pairs, actual *and* possible. This variant of Frege's analysis isn't open to the objection just made.

3.0 Incompleteness[131]

Logicism was a failure. There is nothing wrong with the (set-theoretic) analyses of number, and of number statements, put forth by Frege and Russell and other logicists. What turned out to be false was Frege's claim that there is some *one* algorithm such that, given *any* statement of arithmetic, that algorithm can correctly say whether or not that statement is true or false.

This must be understood aright. Given any finite set of arithmetical

statements, there *is* an algorithm that, given any statement falling into that set, correctly says whether or not that statement is true. But there is no one algorithm that, given any arithmetical statement, correctly says whether or not that statement is true.

One must master some exceedingly intricate mathematics to understand precisely why this claim is false. But some of the philosophical conceits underlying this mathematics are pretty straightforward.

First of all, the word "logic" is ambiguous. Sometimes it refers to ***formal truths,*** e.g., "either Smith is tall or it is not the case that Smith is tall"; and sometimes the word "logic" refers to ***analytic truths,*** e.g., "any case of knowledge is a case of justified belief, but not *vice versa*".[132]

When logicists proclaimed that arithmetical truths are "logical" truths, they meant that they are formal truths.[133] Formal truth, it must be emphasized, is a property of *sentences* and, therefore, of expressions; it is not a property of sentence-*meanings*.[134]

For argument's sake, let's make the following three concessions to the logicist. First, "1 + 1 = 2" is a formal truth in disguise. Second, the logicist paraphrase of "1 + 1 = 2" is a formal truth. We'll refer to that paraphrase as TPT.[135](For the reasons given in section 2.5.1, TPT is some statement along the lines of "two non-overlapping unit classes form a dual class.") Third, what we just said about
"1 + 1 = 2" is true *mutatis mutandis* of any given arithmetical truth.

Bearing these assumptions in mind, consider the statement that:

(EQ) "The logicist translation of '1 + 1 = 2' is TPT."

EQ is not a formal truth. After all, it has the same form as:

(EQ_2) "The Spanish translation of 'there will be justice' is TPT,"

which is obviously false.

It may be that EQ is *analytically* true. But that fact hurts logicism as much as it helps it. To describe a sentence as "analytically true" is to make an elliptical statement about a proposition. It isn't analytic that "triangles have three sides" is true. That sentence could have meant anything; it's an

empirical fact that it means what it does and, therefore, that it's true. What *is* analytic is the *proposition* that, the semantic rules of English being what they are, "triangles have three sides" must be true. The same thing *mutatis mutandis* is true of any other "analytic" sentence, including EQ.

As previously stated, all formal truths are analytic truths (but not *vice versa*). Thus, to describe a sentence as "formally" true is to make a statement about the proposition describing the relationship between the semantic rules of the language to which that sentence belongs, on the one hand, and that sentence, on the other: it is to describe that proposition as analytic.

There is thus no way to ***formalize*** every sentence describing such a proposition. If S_1 is any one such sentence, formalizing S_1 involves using some other sentence S_2 that describes the relationship between the semantic rules of the language to which S_1 belongs, on the one hand, and S_1, on the other. S_2 will invariably fail to be a formal truth. This is because S_2 will have the form: ‹given such and such semantic rules, it follows that S_1 is formally true›, which isn't formally true. The reason it isn't formally true is that it has the same form as the false sentence: ‹given such and such semantic rules, it follows that "grass is green" is formally true. ›

Thus no matter how many truths one formalizes, the justifications for those formalizations will always be given by informal truths. This means that, given any viable method M of formalizing some body of informal analytic truths, there is no complete formalization of the truths that establish the viability of M.

3.1 The concept of a formal procedure

In the last section, we often used words like "formal" and "formalize." And in what follows, we'll often use the term "formal procedure." We'll also use the related term "decision-procedure." So let us make it clear what these terms mean.

Consider the statement: "2 + 2 = 4." You know that it's true. But *how* do you know this? You know what it *means;* and you know that, given what it means, it must be true. Your judgment was based, not on the form, but on the

content, of that sentence.[136] Your ***decision-procedure*** (your way of determining whether "2 + 2 = 4" is true or not) wasn't formal.

Your decision *would* have been a formal one *if* it had been based entirely on rules that had nothing to do with the meanings of symbols. A story will clarify this statement. Although you are intelligent and literate (in your native language), you don't speak English and you don't know what "1", "2", "+", etc. mean. (You are a good mathematician; but in your native country, different symbolic conventions are used to express mathematical statements.) You are on a military mission. This mission involves your being able, within certain very narrow limits, to distinguish between false arithmetical statements and true ones. Your commanding officer tells you that an inscription whose initial segment is "2 + 2 =" is true if it ends with a "4" and flase if it ends with a "5." You have no idea what "2 + 2 = 4" means. (You're not supposed to know.) You come across inscriptions of
"2 + 2 = 4" and "2 + 2 = 5"; and, on the basis only of your commanding officer's instructions, you judge that the first is true and the second is false. In *this* context, your decision-procedure was a formal one.

This decision-procedure had minimal scope. Obviously mathematicians want to find decision-procedures that deal with large classes of statements. In order for there to be any decision-procedure for a sizeable class of arithmetical truths, those truths *cannot* be expressed in the customary way. Supposing the standard symbolic conventions to be in place, the truths of arithmetic *cannot* be expressed using symbols like "2", "+", etc. The reason is that, when arithmetical propositions are expressed in that way, there is no non-semantic method whereby true arithmetical sentences can be distinguished from false ones. Thus, in order for there to be a decision-procedure for arithmetic, sentences such as "2+2=4" must be rephrased (or "paraphrased," as we'll say). Logicists attempted to provide the requisite paraphrases. And, thanks to their paraphrases, it was possible to identify formal-decision procedures for huge classes of arithmetical truths.

Kurt Gödel (1906–1978) showed that there is no way of paraphrasing or otherwise representing arithmetical truths that permits the construction of some one formal decision-procedure that decides the truth of *any* given arithmetical truth. He showed that logicism was doomed and also that *any*

other attempt to formalize arithmetic was doomed. Let us now discuss his achievement in a little (operative word "little") more detail.

3.2 Incompleteness (continued)

Gödel mathematically proved that there is no one formal decision-procedure such that, given any arithmetical statement, that procedure correctly says whether or not that statement is correct.[137] His proof is extremely complex. But here is a way of stating the main idea that, although both inaccurate and incomplete, will at least initiate an interest on the part of the reader in the fascinating discipline of mathematical logic.

First a definition: An "arithmetical" predicate is one that expresses a property of whole numbers, or a relation holding among whole numbers, that can be understood entirely in terms of (i) the concepts of addition and multiplication; (ii) the concepts expressed by the connectives "or", "not", "and", "for any" ("all"), and "for some" ("there exists"); and (iii) the numbers zero, one, two, etc. (i.e., the members of the least inclusive set that contains zero and the successor of anything that it contains). Examples of arithmetical predicates are "is even", "is less than five", and "when added to seven yields twelve." (To say that 3 is less than 5 is to say that *there is* some *whole number* n such that, when n is *added to* 3, the result is 5.)

Given any arithmetical predicate P, there is a class of numbers C such that a whole number n belongs to C iff P(n) (read: "n has P"). Consider, for example, the predicate "is even." There is obviously a class containing every number n such that ‹n is even› is true and containing no number m such that ‹m is even› is false.

For argument's sake, suppose there to be formal procedure FM such that, for any whole number n and any arithmetical predicate P, FM assigns a TRUE to the sentence P(n) if that sentence is true, and assigns a FALSE to if it's false. (In this context, TRUE and FALSE may be thought of as physical stamps. When given an ink-deposit (or whatnot) that expresses an arithmetical proposition, FM stamps TRUE on that ink-deposit if it's true and a FALSE on it if it's false.) This is the same as assuming that:

(1) There exists a formal procedure that correctly says of any given arithmetical statement whether it is true or false.

Any given arithmetical predicate contains a finite number of letters. There are only finitely many different letters. This means that it's possible to produce a list such that, given any finitely long combination of letters Z, there is some whole number n such that Z is the nth member of that list. *A fortiori* it's possible to produce a list such that, given any arithmetical predicate P, there is some whole number n such that P is the nth member of that list. Let L be such a list; and assume that the items composing L are arranged alphabetically. Further, let L* be a list whose first entry is 0, whose second entry is 1, and so on.

Given either list, the nth item on it can be thought of as being "coordinated" with the number n, i.e., as having n for its coordinate. Thus, the first item on L* (namely 0) has coordinate 0; the second (namely 1) has coordinate 1; in general, the nth item on L* has coordinate $n+1$. To simplify exposition, let's suppose that "is bigger than ten," "is even," and "is less than twenty" are, respectively, the tenth, twentieth, and thirtieth items on L.

Let G be a graph whose x-axis consists of L* and whose y-axis consists of L. Thus, the coordinates on G of the statements "1 is bigger than 10," "2 is even," and "3 is less than 20" are, respectively, (2,10), (3,20), and (4,30).

FM assigns FALSE to "1 is bigger than 10" and a TRUE to "2 is even" and to "3 is less than 20." FM can thus be thought of as stamping a FALSE on (2,10), a TRUE on (3,20), and a TRUE on (4,30).

Let K be a class such that, for any whole number n, n is a member of K iff FM assigns a FALSE to (n,n). And let Q be a predicate such that Q(n) is true iff n is a member of K. Supposing that Q is an arithmetical predicate, it follows that Q has some coordinate q on the y-axis. It also follows that the coordinates of Q(q) are (q, q) and, consequently, that Q(q) is true iff Q(q) is false.

Q is an arithmetical predicate. Let's take a moment to make it clear why this is so.

First, for any numbers m and n, ‹m has position n on the x-axis› is equivalent with ‹m=n›, which is obviously an arithmetical statement.

Much the same holds of ‹m has position n on the y-axis›. The predicates composing the y-axis are arranged alphabetically. Each letter of the alphabet can be coordinated with a number. ("A" can be coordinated with "0"; "B" can

be coordinated with “1”; and so on.) So the order in which they occur is to be understood *strictly* in terms of such garden-variety arithmetical facts like “1 is less than 2” and “2 is less than 3.” A consequence is that, for any arithmetical predicate m and any number n, ‹m has position n on the y-axis› is equivalent with some arithmetical statement.

Thus, given either axis, an item m’s having position n on that axis is a strictly arithmetical fact. It follows that, given any arithmetical statement S and any whole numbers m and n, the statement ‹S has coordinates (m,n)› is equivalent with some garden-variety arithmetical statement. A consequence is that, for any statement S and any whole numbers m and n, the statement ‹FM stamps S with a FALSE› is equivalent with some purely arithmetical statement. Since, for any n, Q(n) is equivalent with ‹FM stamps the statement with coordinates (n,n) with a FALSE›, it follows that Q(n) is an arithmetical statement and, consequently, that Q is an arithmetical predicate. Thus, Q(q) is an arithmetical statement. Since, as we saw, Q(q) is true iff it’s false, it’s false. (No true statement entails its own negation.) But FM stamps a TRUE on it, even though, by supposition, FM stamps a TRUE only true statements.

We started with the assumption that there exists a procedure that correctly says of any given arithmetical statement whether it is true or false. This was the only substantive (as opposed to purely procedural) assumption that we made. This assumption entails a false statement and is therefore itself false. There is thus no formal decision-procedure for arithmetic. There is no correct and complete formal characterization of arithmetical truth. Logicism, the doctrine that arithmetical truth is formal truth, is therefore false.

4.0 Analysis by abstraction[138]

In the process of generating his insightful analysis of number, Frege used, for the first time in history, a powerful technique known as ***definition by abstraction.***

The term “definition by abstraction” refers to a way of analyzing properties. This term is thus a misnomer. Expressions are defined; properties aren’t expressions. I’ll therefore drop the term “definition by abstraction” and use the term “analysis by abstraction” in its stead.

Analysis by abstraction is a technique for analyzing comparative properties. (A comparative property is one that can be had to varying degrees.

Thus, weight and speed are comparative properties, whereas four-sidedness is not.) Given a comparative property phi, an analysis by abstraction of phi is given by a statement that explains what phi is in terms of what it is for one thing to have phi to a degree equal to, or exceeding, the degree to which some other thing has phi.

It can be said what it is for two sets to have the same number of members, or for one set to have more members than another, without knowing how many members either set has. If the members of the one set can be paired off with the members of the other, they have the same number of members. If not, the set whose members don't all have partners is the better populated one. Thus, for two sets to be similar (capable of being paired off in the way just described) *is* for them to be equally populated. We may therefore *identify* the degree to which a given set is populated with the class of (possible[139]) sets that are similar to it, and we may also identify a set's being populated to that degree with its being a member of that class.

This technique can be used to analyze any comparative property. (In fact, there is no other way to analyze such a property.) To say that one event preceded some other is to say that the second exceeds the first in respect of the lateness of the time at which it occurred. One can know that one event preceded some other without knowing when either occurred. If there is some possible causal process (e.g., a light ray), beginning with the one and ending with the other, the event on the receiving end of that process is the later one; if there is no such process, the two events are simultaneous. Let us say that one event is "indifferent" to another if there is no possible causal process beginning with either and ending with the other. (If x is indifferent to y, y is indifferent to x.) Two events are simultaneous just in case they occur at the same time, and they are simultaneous if either is indifferent to the other. We may therefore *identify* the time at which an event occurs with the class of events that are indifferent to it.[140]

We defined ‹x precedes y› as ‹some causal process *begins* with x and *ends* with y.› Since the italicized terms are obviously temporal in nature, our analysis might be thought to be viciously circular.

This isn't so. Our analysis appears circular *only* to the extent that it is question-beggingly assumed to be false. If that analysis is right, the concept

of causal-influence (i.e., of one's event being the initiator of a causal process of which the other is the recipient) is *more primitive* than that of order in space-time. So the concept expressed by ‹x causally influences y› (in other words, ‹x is the initiator of a causal process of which y is the recipient›) is the primitive concept, and the concept expressed by ‹x precedes y› is the derivative one.

It's true that we defined ‹x influences y› as ‹some causal process *begins* with x and *ends* with y›. But that was merely a way of *identifying* the meaning of that expression. That act of identification, being one that preceded the analysis in question, had to be given in terms that would be meaningful to those who still held onto the idea that causal influence was to be understood in terms of order in time, and not *vice versa.*

The concept of weight is to be understood along similar lines. One needn't know how heavy two objects are to know that they have the same weight or that the one is heavier than the other. That means that, given any two objects, there is some relation such that their having the same mass can be non-circularly identified with the one's bearing that relationship to the other. An object's weight (the degree to which it has weight)

may therefore be identified with the class of all things to which it bears that relationship. The same thing *mutatis mutandis* holds of speed, length, conductivity, and of any other given comparative property.

A moment ago we said that one can know that x is heavier than y without knowing the weights of either. This must be qualified. When you say that you know only the *comparative* weights of x and y, but don't know the *actual* weight of either, what you are really saying is that, although you know how x and y compare to each other in respect of weight, you don't know how either compares in that respect to objects in general. And when you say that you know *the* weights of x and y, and not merely their relative weights, you are saying that you know how each compares in respect of weight, not only with the other, but with objects in general.

Thus, *all* knowledge of weight is comparative knowledge. The difference between knowing how x's weight compares with y's, on the one hand, and knowing how many pounds x weighs, on the other, is simply one of degree. ‹x weighs 27 lbs› is a condensed way of reporting a great many *relative* judgments concerning x's weight—of reporting a great many judgments concerning x's weight *relative* to some other object's weight. What we think of as non-comparative knowledge is a condensed mass of comparative knowledge. It follows—though I leave it to the reader as an exercise to say exactly *how* it follows—that no viable analysis of a weight, or indeed of any other comparative property, is ever *not* an analysis by abstraction.

5.0 Putting Frege's successes into perspective

Even though logicism turned out to be a failure, Frege's formalization of logic was a success. (This is subject to heavy qualifications that lie outside the scope of the present work.[141]) Many have thought that, to the extent that he was able to formalize logic, Frege succeeded in "mechanizing" the making of inferences—in showing that no thought need be involved in the making of inferences and, therefore, that inferences could be made by beings (e.g., punch-card automata) that are otherwise incapable of thought.[142] The idea is that, given a sentence like:

(1) "for any object x, if x is a mammal, then, if x is intelligent, x doesn't have

pointy ears; and Smith is a mammal and Smith has pointy ears,"

a purely morphology-driven entity—i.e., an entity, e.g., a scanning-device that had no understanding of what anything meant and whose activities are driven only by the *shapes* of the symbol-tokens involved—could infer that:

(2) "it is not the case that Smith is intelligent."

It's granted, of course, that Frege formalized only a miniscule class of inferences. But it's held that *to the extent,* however limited that extent was, that he was able to formalize the making of inferences, he showed that suitably constructed machines could think. And those who hold this hold that others who, like Frege, have formalized some class of inferences have *ipso facto* shown that inferences could be made without thinking and could therefore be made by non-thinkers.

This is deeply absurd. First of all, it's tantamount to saying that thinking can be done without thinking. To make an inference *is* to think. So an unthinking being *ipso facto* can't make inferences.

This argument is likely to be seen as nothing more than a joke. And it is a joke. But it's a joke only because, like any good joke, it states the obvious.

In any case, here's a joke-free argument. Suppose that device M is given tokens of (i) and (ii) as inputs and that, in response, it outputs a token of (iii). It doesn't follow M has *inferred* anything. If M's outputting T_3 is to embody an inference of any kind, it isn't enough that T_1 and T_2 be what led M to do so. It is necessary[143] that their being *symbols* be what led M to output T_3.

Imagine the following. You say to me: "The one copy of your manuscript has just been destroyed." As a result of your saying this I faint. But I faint *not* because of what your utterance meant. Rather, I faint because you spoke so loudly that it traumatized me. My response wasn't of a linguistic nature. It's true that it was caused

by a linguistic act on your part. But your act's *being* a linguistic act was irrelevant; it could just as well have been an explosion or the ringing of a bell.

Let's change the story. Because of experiments that I underwent as a

child, I say “four” whenever I hear very loud sounds. You ask me “what is 2 + 2?” and I say “four.” But I say it *not* because of what it was that you were asking, but because you said it so loudly that my conditioning kicked in and I reflexively barked out “four.” My saying “four” was caused by a linguistic act on your part. But it wasn’t caused by your act’s *being* linguistic in nature. For this reason, my response wasn’t linguistic in nature.

For exactly similar reasons, given only that M is outputting T_3 in response to its being fed inputs T_1 and T_2, it doesn’t follow that M’s outputting T_3 embodies an inference. Since it wasn’t *in virtue of* their being symbol-tokens that T_1 and T_2 led to M’s outputting T_3, M’s outputting T_3 is no more expressive or constitutive of a linguistic or inferential or otherwise ratiocinative act on M’s part than is my inadvertently getting a nose-bleed as a result of my hearing an extremely loud (and therefore nose-bleed-inducing) noise that happens also to be a case of somebody’s saying “if you get a nose-bleed right now, I’ll give you a million dollars.”

Thus, so far as it’s correct to say that Frege “mechanized” thought, what it means is not that, thanks to his efforts, *no* thought is needed to execute inferences that were previously thought-dependent. It means that inferences that, prior to his efforts, had to made *ad hoc* could now be made in a systematic manner. Adding two multi-digit numbers together is easy if you use the standard algorithm, and it’s hard if you don’t. But, even if you use it, thought is involved. *Applying* the algorithm takes thought; and *knowing that it’s applicable* takes thought. Only a thinking being can make the judgment that given certain ink-marks (e.g. “456 + 785”) it is to be used, whereas given others, it isn’t. So even though it’s a lot harder to add multidigit numbers without the algorithm than with it, adding them with it involves thought.

6.1 Putting Frege’s successes into perspective (continued)

The inferences one must make *in order* to algorithmize a given task are often harder to make than the inferences that the algorithm, once applicable, will enable one to make. The same is true of the inferences one must make in order to superintend the application of an already identified algorithm.

Once again consider Arguments #1 and #2. They're obviously valid and invalid, respectively. True—thanks to the algorithms bequeathed to us by Frege, Argument $\#1_{RP}$ is easier to evaluate than Argument #1. But what it takes in the way of intelligence to convert Argument #1 into something to which those algorithms are applicable is comparable to, if it doesn't exceed, what it takes to decide whether or not Argument #1 goes through.

Also, it's no easy matter to figure out *whether* a given statement is expressed in a form suitable for the application of one of those algorithms. There are, as we know, arguments superficially similar to #1 that aren't valid, and it's only through the use of judgment that we can distinguish such impostors from their legitimate brethren. There is no mechanical way to do so. In general, there's no mechanical procedure for determining the validity or the scope of one's mechanical procedures, a consequence being that thought is inherently incapable of being mechanized.[144]

Imagine the following. You run a company that makes cars. You recently hired a new worker, who goes by the name of "Bucky." Unlike your current employees, Bucky isn't intelligent or skilled. But he works hard and he works cheap; and he's an extremely good worker *once he's mastered a given task.* However, Bucky can only do rote, mechanical tasks; and the amount of training he must undergo *before* he can perform a given task is twice what your average employee would have to undergo to perform that same task. Also, Bucky requires an unusual amount of supervision. He's unusually likely to break equipment; and, being psychologically fragile, if conditions in the factory are changed slightly (e.g., the walls are repainted), he becomes confused and has to be coddled for a while before he can again work effectively.

The benefits of keeping Bucky on will exceed the costs if, and *only* if, there is a large amount of rote work for him to do *and* you can spare the man-power needed to train and supervise him *and* you know that conditions won't change in the factory in such a way as to render him incapacitated—and so on.

Algorithms are a lot like Bucky. A lot of time and effort is needed to put them to work; and once put to work, they can only do very specific things. If you need those specific things to be done often enough, it's worth it to

develop an algorithm for doing them and it's worth it to check in from time to time to make sure that they're being applied properly. But they're useless unless heavily supervised, and they're thus no substitute for real thought.

Chapter 8

The Relationship between Thought and Language:

Literal Meaning vs. Cognitive Content

1.0 The objects of thought are always truths or falsehoods, and never objects, except in a derivative sense

We obviously think about rocks, trees, people, and other things. But things are the objects of thought only in a derivative sense. The real object of thought is always some truth or falsehood *concerning* a thing.

Some fiction will help. At time t, you are looking at a rock. You recognize the rock. It's the rock that Joey, your arch nemesis, threw at your car. That incident was an emotionally charged one for you, and you decided to write a book about it. In order to have a concise way of referring to it, you gave it a name: you decided to call it "R." It's been a while since you've seen R. But when, at time t, you see it, you instantly recognize it. You recognize it as R—as the rock that Joey threw at your car.

Your eyes are obviously telling you something. But what exactly? Here we must be careful to distinguish what your visual perception *itself* is telling from what, given the backlog of knowledge at your disposal, you can read into that perception. No sooner do you see R than you *recognize* it. But do your eyes tell you that it is R you are seeing? No. Given only what they are telling you, you could be looking at some other rock that happened to look just like that rock. Maybe you know that this is unlikely. Maybe you know that there's basically *no* real chance of this. And maybe you know all of this instantaneously. You don't have go through a conscious, drawn out process

of inference. It's obvious—just as obvious as it is that the person in front of you at your wedding is your beloved fiancée.

But *your eyes aren't telling you any of this.* Given only the information encoded in your visual perception, you *could* be seeing some duplicate of R. Given only what your eyes are telling you, it's possible that R no longer exists and that what you are seeing is a perfect duplicate of it.

What your eyes tell you isn't: "that's R over there." What they tell you is:

(RO) "there is a rock having such and such characteristics over there."

R *satisfies* that existence-claim; and you know that. On the basis of that knowledge *along* with the information bequeathed to you by your visual perception, you know that R has such and such characteristics. But that perception itself doesn't bear any R-specific message. What your eyes are telling you, taken by itself, is consistent with R's not even existing. Obvious extensions of these points show that, given any message given to you through vision or any other sensory modality, and given any object in the external world, that message is consistent with that object's not existing.

1.1 The exact way in which sense-perception is descriptive

Our sense-perceptions obviously represent objects to us and can thus be said to describe them. But what isn't obvious is that *there is nothing to perceiving anything other than being perceptually apprised of some existence-claim satisfied by that object.*

Let's suppose you are looking at your good friend Pat. Your seeing Pat obviously involves your seeing instances of various properties of various different kinds. Some of these property-instances are morphological (shape-related); others are chromatic (color-related); others are kinematic (motion-related). And so on.

But you don't *just* see various property-instances. You see an organized structure consisting of them. Your visual perception of Pat represents these property-instances as bearing various relations to one another. Being thus interrelated, they collectively form a coherent structure. Your seeing a coherent structure of property-instances, as opposed to a disorganized jumble

of them, consists in your seeing various instances of various relations. It consists in your seeing that certain property-instances are to the right of certain others, that other property-instances are behind certain others, etc. Thus, your seeing Pat involves your seeing various instances of various properties and relations.

It follows that your seeing him involves your being (visually, not verbally) apprised of an existence-claim: "there are, at the present moment, instances of such and such properties, in such and such places, bearing such and such relations to one another." From now on, when I mention "property instances," I am referring to instances of properties *and* of relations.[145]

We must now confront a delicate but deeply important question. You see Pat. That's a given. This *involves* your seeing various instances of various properties and relations. That too is a given. But do you see anything *other* than property-instances instances? Is there something to your seeing Pat above and beyond your beholding these various instances? Do you see these instances, on the one hand, *and* Pat, on the other?

No. Clearly you just see the instances. Your perception doesn't represent Pat *in addition to* the various property-instances that are associated with him. Were your perception to have such a structure, it would involve Pat's being represented to you as a featureless lump, a blank undifferentiated entity, with various property-instances piled on top of it, so to speak.

But that clearly isn't how that perception works. It's absurd to suppose that a thing possibly be perceptually represented *independently* of the property-instances that it is being represented as having. That would involve its being represented as a featureless blank. The very idea of a thing's being perceptually represented *as* a featureless blank is a non-starter. Thus, there is nothing to seeing (or otherwise sense-perceiving) an object other than seeing (same qualification) the property-instances associated with it.

These properties are time- and place-indexed. Supposing what you are seeing is round, what you see isn't just an instance of roundness, but an instance of roundness *in a certain place, at the present time.* So if O is what you are seeing, your seeing O consists in your being visually told that there are instances of various (time- and place-indexed) properties, these being properties had by O.

Your seeing a given object consists in your being apprised of the truth of an existence-claim that it *uniquely* satisfies. And your seeing that a given object has property P consists in your being apprised of the truth of some existence-claim of the form: there exists exactly one object having such and such properties; moreover, any such object has P.

Let me clarify these statements. Look at the last word of this very sentence. Your eyes don't *only* tell you that certain (time- and place-indexed) properties are instantiated. They don't *only* tell you that, at the present time, in a certain place, there is an instance of a certain morphology (i.e., that had by the following inscription: "sentence") and of certain chromatic properties, and so on. They tell you that, in that place, at that time, there is exactly *one* such instance.

Thus, you see objects by being visually apprised of existence-claims that they *uniquely* satisfy. And, of course, what we just said about vision is true of all sensory modalities. Perception is purely descriptive. For you to perceive an object is for you to be apprised (in a certain way) of an existence-claim that it uniquely satisfies. (Perceiving an object involves being causally connected to it. Some think this is inconsistent with the idea that perception is fundamentally descriptive in nature. Reasons to think otherwise are given later in the present Chapter.)

1.2 The pre-semantic basis of communicated meaning

The idea that anything should be perceptually represented as a featureless blank is not a complete non-starter, and so is the idea that anything should be *cognitively* represented in that way. *Con*ception, like *per*ception, is and must be fundamentally descriptive. To think of an object is to think some existence-claim that it uniquely satisfies. Russell (1917) saw this.

But it's easy to confuse issues relating to conception with issues relating to language. What people think isn't always in complete lock-step with what is literally meant by the expressions that they use. Cognitive content doesn't always coincide with literal meaning. There is *a* systematic relationship between the two—but it isn't always one of identity.

Another example will make it clear what I mean by this and will help us avoid confusing facts about thought with facts about language. (And this, in its turn, will help us develop a correct semantics *along* with a correct understanding of thought.)

One day, you meet a person—you've never met this person before. You shake hands and introduce yourselves. He tells you his name ("Fred"). Of course, you see what he looks like and how, at that particular moment, he is behaving. You see that he has such and such features (e.g., he has brown hair), is moving about in such and such manner (he's nervously wringing his hands), and wearing certain clothes (e.g., a checkered shirt).

It is clear that, in seeing Fred, you are being a given a *description* of him. Your perception *describes* him to you as having such and such characteristics. In fact, it's clear that your seeing him *consists* in your being given a description of him—not a verbal description, of course, but a visual description. Thus, your seeing him consists in your being a visually encoded description of Fred, and your seeing him for any length of time consists in your being given a series of such descriptions.

Your visual perception tells you that, standing in front of you, at the present moment, there is a person (or, in any case, person-like entity) having such and such features (e.g., it has a certain general shape and height; it has certain facial and bodily characteristics; it has certain kinematic (movement-related) characteristics; it's wearing clothes of a certain kind; and so on). Given that this person has told you his name, the information communicated to by your senses is, when described very approximately and abstractly, along the lines of[146]:

(F_1) There is, at the present time, in a certain place (relative to what where I am located at the present time) an object x having such and such characteristics (e.g., its morphology is characteristic of a human being, it's wearing a checkered shirt, it's nervously wringing its hands, etc.); and, given what x just said, it can safely be assumed that "*Fred*" *refers to x.*

The italicized part identifies the semantic rule for "Fred." Let's refer to that rule as SF (this abbreviation will be frequently used in what follows). Notice how descriptively impoverished SF is. All it does is assign a certain

label to a certain object. F_1, by contrast, is a descriptively very rich claim; it says a lot about the world. And you learn SF *through* F_1.

Semantic rules can't just be known. They must known through sense-perception. "Socrates" refers to Socrates. But we could refer to him as "Gronk" if we chose to, and it's only because you and I have had certain sense-perceptions that we know that "Socrates," not "Gronk," is his name.

But the sense-perceptions through which we learn semantic rules are replete with information that has nothing to do with sense-perceptions. The semantic information is always embedded in non-semantic information. Your learning SF is embedded in your learning F_1, for example.

Because of this, what sentence-utterances *convey* may contain vast sums of information that they do not have for their literal meanings. Let's suppose that, a few hours after you meet Fred, your friend Larry produces the following utterance while addressing you:

(LS) "Fred is a professor of economics."

LS attributes a certain property (that of being an economics professor) to a certain individual. And that's all it does. It's a very simple statement. If that individual has that property, then LS is true; otherwise, it isn't. To put it more formally, there is some individual x (namely, Fred) such that LS attributes the property of being a professor of economics to x and such that LS is true if and only if

(LP) x is a professor of economics.

But given how you *learned* who "Fred" refers to, what LS conveys won't be so threadbare. True—for some object x (namely Fred), SF merely slaps the label "Fred" on x. But your *knowledge* of SF is replete with non-semantic content. You know of Fred's existence only because he was *described* to you. In seeing him, you were visually told that there was some individual x such that you met x at a certain time, in a certain place, and such that x had certain characteristics; and you were also told that "Fred" was x's name. Your perception of him was therefore descriptive, and the same is therefore true of your subsequent *concept* of him. Given that you cognitively access SF through this information, what LS communicates to you is likely to be along

the lines of:

(LPC) There was, a little while ago a certain object x having such and such characteristics (e.g., was wearing a checkered shirt and nervously wringing its hands, etc.) and "Fred" refers to x and I've just been told that: *x is a professor of economics.*

It should be pointed out that you'll know that the *literal meaning* of LS is confined to the italicized part. You'll know that "Fred" isn't *synonymous* with ‹the guy who I met at time t and place p, who was wearing a checkered shirt and wringing his hands nervously [etc.].› And you'll know that does not, at the level of literal meaning, say anything about anyone's ever wearing a plaid shirt or wringing their hands.

At the same time, given the circumstances under which you learned who "Fred" refers to, it's inevitable that LS would convey some such message to you. Given the circumstances, the communicated, as opposed to literally meant, message soaked up some of the non-literal, descriptive information implicated in your initial tête-à-tête with Fred.

To sum up, there is some individual x such that, in terms of its literal meaning, all LS does is to attribute the property of being an economics professor to x. So what LS literally means is pretty spare. But what it incidentally conveys to you isn't so spare.

1.3 Interlude: Mill on reference

A moment ago we said that, for some object x (namely Fred), SF merely slaps the label "Fred" on x. This point, appropriately generalized, is that, given any proper name N, there is some object O such that N's sole linguistic function is pick out O—such that, in other words, the semantic rule for N is simply: "N refers to O." Names "denote" (pick out) but do not "connote" (describe). They are mere labels.

This is what we said, and it's plausible. But it's not exactly self-evident. So we need an argument for it. John Stuart Mill provided that argument.[147]

Imagine, said Mill, that you decide to refer to a certain river as the "Dartmouth," the reason being that, at the time of your naming it, that river's mouth was at the sea-side city of Dart. (Let's suppose that you're royalty, so

that your linguistic practices automatically become part of the English language.) Should that river change its course as a result of some cataclysm, so that its mouth were no longer anywhere near the city of Dart, it would still be called "the Dartmouth." And therein lies the distinction between "Dartmouth," which is a proper name, and "the river whose mouth is at the city of Dart," which is not a proper name. The first merely labels a certain river; it doesn't pick it out *through* a description. That's why it labels that river *even after* that river ceases to fit the relevant description. And that's why, if some other river were to change locations, in such a way that *its* mouth coincided with the city of Dart, "the Dartmouth" wouldn't refer to it, and would still refer to the other river (the one whose mouth was originally at Dart).

By contrast, "the river whose mouth is at the city of Dart" is not a mere label. (Like the authors whose work we're about to evaluate, we are in this context operating on the assumption, definite descriptions are referring terms.) It refers to whatever it is that fits a certain *description* (i.e., "unique river whose mouth is at the city of Dart"). So it refers by describing, not by labeling.

Here is another, similar argument. Suppose that You and I are spies in enemy terrain. We see a person in the distance. We can't make out who it is. We don't know who he is. But, the circumstances being what they are, we know that, whoever he is, he's important. You and I are going to be talking about him a lot. We thus need a way of referring to him that, unlike the "the man in the distance," won't cease to be useful the moment he moves to a new place. We thus decide to give him a name. We name him "Gargantuan."

Russell and Frege would say that "Gargantuan" is synonymous with "the man in the distance" or, at any rate, with some other definite description. But we've already seen why that view is the wrong one. That person, whoever he is, can move to a new place; and some other person who isn't identical with him can move into the place that he has vacated. But "Gargantuan" would not on this account cease to refer to one person and come to refer to someone else. Let us develop this point by telling a story.

We finally meet Gargantuan. He turns out to be Mr. X, a high-level official in the government of the nation we're spying on. We now know X's name, of course. But when speaking to each other, we still refer to X as

“Gargantuan.” We befriend X, and he turns out to an amicable and easy-going fellow. He is no longer distant in any sense, and “the man in the distance” no longer refers to him. But “Gargantuan” *does* refer to him. “The man in the distance” picks out whoever it is that fits the relevant description, and that person won’t always be X. But Garguantuan always picks out X. Given any property phi, such that, at some point in time or other, ‹the phi› picks out X, a similar argument shows that there are possible circumstances where ‹the phi› fails to refer to X, even though, in those same circumstances, “Gargantuan” still does refer to X.

Thus, “Gargantuan” doesn’t refer to X by describing him, and X’s being picked out by “Gargantuan” isn’t contingent on his having this or that property. “Gargantuan” simply labels X. All X has to do to be labeled by “Gargantuan” is continue to be X. By contrast, to be referred to as “the man in the distance,” one has to be a man in the distance, which isn’t always easy.

Mill’s argument decisively shows that proper names are mere labels. They are not “connotative;” they don’t have Fregean “senses.” By contrast, definite descriptions (supposing them to be referring terms) *are* connotative; they *do* have Fregean senses.

Non-connotative referring terms are said to refer *directly;* connotative referring terms are said to refer *indirectly.* In Chapter 6, it is shown that all reference is direct reference and that the concept of indirect reference is a completely incoherent one. But we will not assume that here, since the authors whose work we're evaluating don't assume it.

1.4 Perception, conception, literal meaning

The just-discussed principles just are crucial. True—they’re obvious and easy to understand. But for some reason some leading analytic philosophers have overlooked them. The results have been catastrophic. Talented thinkers have wasted decades of their lives defending broken ideas.

To prevent you from suffering the same hideous fate, I’m going to subject you to two more stories that, although little more than boring repetitions of the already boring stories already boringly told, will ensure that you base your philosophical career on the right principles. Somebody with a handle-bar mustache who is wearing a bolo tie who introduces himself to you

as “Larry.” You obviously know that “Larry” isn’t synonymous with “the guy with the handle-bar mustache who is wearing a bolo tie who is introducing himself to me right now.” (When your spouse—who we’ll call “Pat”—was first introduced to you, did you think that “Pat” was synonymous with “the well-dressed charming person with whom I’m now having a conversation”? No!)What you know is that there is a certain individual—an individual who, it is true, has a handle-bar mustache, is wearing a bolo tie, and is currently introducing himself to you—and “Larry” refers to that individual. Thus, if it is stated in a formal and perspicuous manner, what you know is[148]:

(LD) There is some individual x such that x uniquely has the property of being a person with a handle-bar mustache who is wearing a bolo tie who is introducing himself to me at this exact moment; and “Larry” refers to x.

Thus, there is some x such that the semantic rule for “Larry” is simply:

(LSR) “Larry” picks out x.

A corollary is that there is some x such that, at the level of literal meaning:

(LT) “Larry is a talented astrologer”

is true just in case x is a talented astrologer. Thus, for some x (such that x is Larry), LT has for its meaning the threadbare proposition:

(LX) x is a talented astrologer.

But given the circumstances under which you *learned* the semantic rule for “Larry,” LT will *convey* the descriptively much richer proposition:

(LT_2) There is currently somebody in front of me who is wearing a bolo tie and has a handle-bar mustache [etc.], and that person is a talented astrologer.

So what LT will have one proposition (LX) for its literal meaning, but will non-literally suggest (or, as philosophers of language say, “implicate”) some richer proposition. This second proposition is (similar to) the one meant by the sentence:

(LT_3) “The person currently standing right in front of me who is wearing a bolo tie and has a handle-bar mustache [etc.], and that person is a talented

astrologer."

1.5 Our ability to distinguish literally meant from presemantically implicated content

The following extension of our Larry-story will ram these ideas home and will also help us move onto a new and important point.

Brown is a friend of yours who knows Larry, and Brown is at the party where you meet Larry. While you are talking with Larry, Brown utters LT to you. For reasons already discussed, there is some x (namely, Larry) such that Brown's utterance will have LX for its literal meaning, but will non-literally impart the richer proposition (LT_2).

But—and this is the just-advertised new (and important) point—you know that LT isn't synonymous with (LT_2) or with any other such sentence. You know (at some level) that nothing having to do with bolo ties or handle-bar mustaches is any part of what is *literally* meant by LT.

I don't mean that, *if you have* some training in semantics, you'll know this. I mean that you'll know it merely by virtue of being a competent and cognitively normal speaker of English—that, merely by virtue of satisfying these minimal conditions, you know (or would readily come to know, if you thought about it) that what is literally meant by Joe's utterance doesn't encode any information about anybody's wearing a bolo tie or having a handle-bar mustache.

Here's why I say this. If asked whether:

(LSC) "Larry does not have a handle-bar mustache"

is self-contradictory, like "squares don't have four sides," as opposed to merely false, like "Barack Obama isn't the president in 2009," you'd know right away that it's merely false. You'd know that, although Larry happens to have such a mustache, the idea of his not having one makes perfectly good sense. And given that you speak English, you know that LSC, though false, *could* be true, unlike "squares don't have four sides."

But if "Larry" were synonymous with "the person who uniquely has the property of being somebody in front of me right now who is wearing a bolo

tie and has a handle-bar mustache," then LSC *would* be self-contradictory, since it would then be synonymous with the definitionally false sentence:

(LSC_2)'the person in front of me right now who is wearing a bolo tie and has a handle-bar mustache does not have a handle-bar mustache."

If LSC were definitionally false, you'd know it. You'd know it for the same reason that you know that "squares do not have four sides" is self-contradictory and, for that reason, doesn't have a meaningful role in discourse. You'd know it because your knowing it is constitutive of your ability to speak English—of your ability to use English expressions in accordance with the appropriate rules.

You can't necessarily articulate this knowledge. But that means little. Given how you speak and what judgments you make, it's clear that you are guided by it and, therefore, that it's no less known to you than your own name.

1.6 No non-descriptive knowledge

But—and this must be emphasized—there is no way that you can *non*-descriptively think about Larry (or anything else). Larry can't *just* be represented to you. He must be represented *as* having these or those properties. Circumstances being what they are, he is represented to you *as* the guy currently right in front of you wearing a bolo tie [etc.].

This doesn't mean that you'll *always* think of Larry this way. You won't. Tomorrow, you'll think of him as the guy who *yesterday* was standing right in front of you and, at that time, had a handle-bar mustache [etc.]. And if you come to be friends with Larry, that description will be replaced by others.

But you'll always think of him descriptively. More precisely, it will always be by virtue of your knowing the truth of some existence-claim uniquely satisfied by Larry that you'll be able to have thoughts about him. Your being able to think about him will always consist in your knowing that something uniquely has P, where P is some property that Larry has.

1.7 Reference-fixing vs. meaning giving[149]

Question: how is it that you can *think of* Larry as "the person in front of me

right now who is wearing a bolo tie and has a handle-bar mustache," and yet know that Larry might *not* have had a handle-bar mustache (or been standing right here or been wearing a bolo tie)?

Simple. Remember what Russell said about definite descriptions. (‹The phi has psi› means ‹something or other uniquely has phi, and any such thing has psi.›) In light of this, consider the sentence:

(LD_1) The guy now standing right in front of me who is wearing a bolo tie and has a handle-bar mustache [etc.,] is named "Larry."

There are two ways to interpret LD_1. Here's the first interpretation, which is the correct one:

(LD_2) Somebody x is uniquely such that x is now standing right in front of me and x has a handle-bar mustache [etc.]; moreover, "Larry" is x's name.

Here's the second interpretation, which is the wrong one:

(LD_3) "Larry" names anyone x uniquely such that x has the property of now standing right in front of me and x is wearing a bolo tie and has a handle-bar mustache [etc.].

LD_3 says that "Larry" is *by definition* a unique handle-bar mustache having object x such that x is standing x right in front of me [etc.]. LD_2 says that "Larry" names the person who, under the circumstances, is a unique handle-bar mustache having object x such that x is standing x right in front of me [etc.]. Both LD_2 and LD_3 use the same description ("unique thing x such that x has a handle-bar mustache [etc.]"). But they use it in different ways. LD_3 says that "Larry" *by definition* picks out something fitting that description. LD_2 says that "Larry" picks out the thing that, as it happens, fits that description. More precisely, LD_2 says that something x fits that description and the semantic rule for "Larry"—the rule that you must know to know how to use "Larry" correctly—is: "Larry" refers to x.

So, as Saul Kripke puts it, LD2 uses that description to *fix the referent* of "Larry," whereas LD_3 uses it to *give its meaning* of "Larry. LD_3 is obviously

wrong, since “Larry” obviously isn’t synonymous with “the guy now standing in front of me right who is wearing a bolo tie,” or anything of the sort; and LD_2 is therefore correct.

Given that LD_3 is correct, there is some x such that, at the level of literal meaning, “Larry” does nothing other than pick out x, and such that, at the level of literal meaning, ‹Larry has psi›, does nothing other than attribute psi to x. But for reasons that we’ve made clear, ‹Larry has psi› has non-literal connotations, to use John Stuart Mill’s expression, that very much affect what it conveys and also, therefore, what it is used to convey.

1.8 Identity statements and modality

Suppose that, a day after meeting Larry, you meet a person who is clearly wearing a disguise. (Let’s not worry about why he is wearing a disguise or why you don’t find it alarming.) That person introduces himself to you as “Oliver.” You and he then proceed to have an amicable conversation. At that point, a series of events occur that parallel those of the previous day. Your friend Brown is in attendance and joins the conversation. Brown knows Oliver; and, addressing you, Brown says:

(OT) “Oliver is a talented astrologer.”

Given what we’ve already said, it’s clear what the consequences are. There is some x such that the semantic rule for “Oliver” is:

(OSR) “Oliver” refers to x.

A corollary is that OT has for its literal meaning the descriptively impoverished proposition:

(OT_1) x is a talented astrologer.

At the same time, given the circumstances under which you learn who “Oliver” refers to, what Brown’s utterance of OT will convey to you (non-literally) will be descriptively richer than OT and will be along the lines of:

(OT_2) There is some individual such that, right now, that person uniquely has the property of being a person wearing a disguise who is standing right in front or me; moreover, that person is a talented astrologer.

But for reasons analogous to those given earlier, you will know (at some level) that OT doesn't have OT_2 or any other comparable proposition for its meaning; you know (at some level) that nothing having to do with disguise-wearing is any part of what is literally meant by OT.

But Oliver *is* Larry. You don't know this. It just happens to be the case. Unbeknownst to you, the person you know as "Larry"—the guy with the handle-bar mustache—is identical with the guy you know as "Oliver"—the guy wearing the mask.

There is some one individual who goes by "Larry" and "Oliver."

It follows that there is some *one* individual x such that the semantic rule for "Oliver" is:

(OSR) "Oliver" refers to x,

and such that the semantic rule for "Larry" is:

(LSR) "Larry" refers to x.

There is thus some one individual x (this being Larry) such that, given that OSR and LSR are the semantic rule for "Oliver" and "Larry" Jones respectively, and such that, consequently:

(OT) "Oliver is a talented astrologer"

means

(OT_1) x is a talented astrologer

and such that

(LT) "Larry is a talented astrologer"

means:

(LX) x is a talented astrologer.

In other words, OT and LT *have the very same meaning.* OT_1 and LT are the same proposition. And that one proposition is what is meant by OT *and* by LT. LT and OT mean the same thing.

But we have a problem. OT and LT *convey* very different things. LT

conveys:

(LT_2) There is currently somebody in front of me who is wearing a bolo tie and has a handle-bar mustache [etc.], and that person is a talented astrologer.

OT conveys the proposition that:

(OT_2) There is some individual such that, right now, that person uniquely has the property of being a person wearing a disguise who is standing right in front or me; moreover, that person is a talented astrologer.

LT_2 and OT_2 are obviously very different propositions. Neither entails the other. Neither even makes the other probable (they're confirmationally independent of each other).

But how can this be? How can LT and OT convey such different propositions, given that they have the very same literal meaning?

We've already explained this. Semantics isn't learned in a vacuum. It's learned empirically, that is, through sense-perceptions, and those sense-perceptions convey a lot of information that has nothing to do with the semantic information that they are also conveying. To access these semantic rules, one must work through this pre-semantic, and non-semantic, information; and the result is that what is communicated by a sentence may be much richer than its literal meaning. We've already explained why, under the circumstances in question, these particular sentences (LT and OT_1) happen to convey these particular propositions (LT_3 and OT_2).

Let's continue our story. Addressing you, your friend Brown says:

(LOI) "Larry Jones is identical with Oliver."

You are surprised. After all, "Oliver" was wearing a mask, and (we may suppose) his demeanor was very different from that of "Larry Jones." Your initial reaction is one of disbelief. But "Oliver" takes off his mask and he is indeed Larry Jones. It also turns out that this person is legally named both "Oliver" and "Larry Jones." (We don't have to worry about the details.)

Let's consider the time *just* before Oliver takes off his mask—the time when you are still skeptical about Brown's claim. Letting psi be any property,

obvious extensions of what we've already said show that, under these circumstances:

(LTP) ‹Larry has psi›

will convey something along the lines of:

(LTP*) yesterday, there was somebody in front of me who was wearing a bolo tie and had a handle-bar mustache [etc.], and that person has psi

and

(OTP) ‹Oliver has psi›

will convey something along the lines of:

(OTP*) There is some individual such that, right now, that person uniquely has the property of wearing a mask while standing right in front of me and having a conversation with me that's been going on for a while; moreover, that person has psi.

It follows that, given the circumstances, LOI will convey much the same thing as the sentence:

(DLOI) "the guy who, yesterday, was standing in front of me and wearing a bolo tie [etc.] is identical with the guy who is currently standing in front of me and wearing a mask [etc.]"

We know from Russell's theory of descriptions that the meaning of DLOI is:

(DLOITD) somebody x uniquely has the property of being a person who, yesterday, was standing in front of me while wearing bolo tie [etc.], and having that property being such a person also uniquely has the property of standing in front of me right now and wearing a mask [etc.].

Of course, DLOITD is *not* a trivial proposition. It does not say of some individual x that x is identical with x. It says something radically different. It says that some one individual uniquely has each of two very different properties.

But remember that names are mere labels, and that, consequently, *there is* some x such that "Oliver" and "Larry Jones" simply label x. It follows that there is some individual x (namely, Oliver/Larry Jones) such that what is

literally meant by LOI is simply:

(LOX) x is identical with x,

which, of course, is utterly trivial. But that's ok, since the non-triviality of what LOI communicates has already been accounted for.

1.9 Frege[150] and Russell[151] on proper names

Though they lived after Mill, Frege and Russell didn't internalize his very good analysis of proper names and chose to replace it with a deeply defective one.

Here was their reasoning. You and a friend are looking at the last celestial body to disappear from the morning sky—your friend tells you that it's called "Hesperus." (You don't know anything about astronomy, but your friend does.) Later that day, just as evening is approaching, you and he are again looking at the heavens. He points to the first celestial body that appears in the evening sky and tells you it's called "Phosphorous."

Given only how they appear to you on the occasions in question, you have no way of knowing whether or not "Hesperus" and "Phosphorous" co-refer. If you had to guess, you'd probably say that they did not, given that the celestial bodies referred to by "Hesperus" and "Phosphorous" respectively were in very different parts of the sky and seemed to have quite different features, so far as you could make out their appearances.

In the late morning, just after being told what "Hesperus" refers to, your friend tells you (correctly):

(H_1) "Hesperus is less massive than the Earth."

In the early evening, just after being told what "Phosphorous" refers to, your friend tells you (correctly):

(P_1) "Hesperus is less massive than the Earth."

Given the circumstances, H_1 and P_1 convey very different propositions to you. H_1 conveys the proposition that would be conveyed to you by:

(MS) "The morning star is less massive than the Earth."

(In this context, take "the morning star" to be an abbreviation for "the last celestial body to disappear from the morning sky" and take "the evening star" in the same way *mutatis mutandis*.)

And for the same reasons *mutatis mutandis*, P_1 conveys the proposition that would be conveyed to you by:

(ES) "The evening star is less massive than the Earth."

Thus, the propositions conveyed by MS and ES are, respectively, to the effect that:

(MSTD) something is uniquely a last celestial body to disappear from the morning sky, and any such thing is less massive than the Earth

and

(ESTD) something is uniquely a first celestial body to appear in the evening sky, and any such thing is less massive than the Earth.

Obviously MSTD and ESTD are different propositions. Neither entails the other. They are confirmationally independent of each other.

Since MS and ES convey MSTD and ESTD, it might seem to follow that MS and ES have MSTD and ESTD for their respective literal meanings. And this is just what Russell and Frege inferred.

The Russell-Frege analysis of proper names is this:

> (RF) If N is a proper name of some object O, that is because N is synonymous with ‹the phi›, for some property phi that O, and O alone, has. Thus N is O's name only because O fits the description encoded in O. Proper names are not mere labels; Mill was wrong. They are descriptive.

(Note: Russell and Frege disagreed about how to analyze ‹the phi.› Frege thought that it was a referring term. Russell thought otherwise. But they both accepted RF, and that's what's important here.)

So "Homer" is, in their view, synonymous with "the blind bard who authored *the Iliad,*" or some such, and "Shakespeare" is synonymous with "the author of Hamlet," or some such.

We've already seen why this theory is wrong and also that Mill, whose

work was known to both Frege and Russell, refuted this theory. But, for whatever reason, they didn't internalize Mill's argument. (That happens. There are many arguments that I've read and that, if I had understood them, I would have seen to refute cherished positions of mine.) And, even though (unbeknownst to them) there was a cogent reason *not* to hold it, there were some excellent reasons *to* hold it. In other words, apart from the fact that it's wrong, it's an excellent theory—one that models the data pretty well.

Let us state these reasons as explicitly as possible. There are basically three.

1.9.1 Reason #1: Since conception is descriptive, and since language mirrors thought, proper names are descriptive

Russell and Frege believed that conception is descriptive—that to think *O has psi* is to think *the phi has psi,* where *phi* is some property had by O and O alone.

They were 100% right about this. For reasons to be discussed in a moment, this insight of theirs, though accepted for the first 70 years of the last century, was promptly rejected in the early 1970s (for not very good reasons).

Here's where they went wrong. They assumed that what is true of language must be true of thought. Thinking (rightly) that one's *concept* of Socrates, if put into words, would have the form ‹the phi›, for some property phi, they inferred (wrongly) that "Socrates," the word, must be synonymous with ‹the phi.›

They're wrong, as we've seen; and, contrary to what they thought, the descriptive nature of conception is perfectly consistent with the supposition that proper names are mere labels.

What made it possible for them to overlook the flaws in their logic was that they focused on names of historical figures ("Homer," "Shakespeare," "Napoleon") or celestial bodies ("Hesperus," "Phosphorous")—names that are likely to be associated with a single definite description. It isn't likely that "Phosphorous" will stop being "the evening star." (It could happen; there could be some cataclysm. But it isn't likely.) And it isn't likely that "Shakespeare" will stop being synonymous with "the author of *Hamlet.*" (It's possible; it could be discovered that Francis Bacon wrote *Hamlet.* But it isn't

likely.) So each of “Hesperus” and “Shakespeare” is tethered to a definite description whose aptness events probably won’t threaten.

But this isn’t true of names of living people. Remember what we said about “Larry.” What makes it so obvious that “Larry” isn’t synonymous with “the man with the handle-bar mustache [etc.]” is that Larry might shave his mustache tomorrow; maybe he already has. And maybe I’m now at some other party and am talking to some other man with a handle-bar mustache, and it is now this other man, not Larry, to whom “the man with the handle-bar mustache [etc.]” corresponds. Since living people change and, as it were, move from definite description to definite description, it’s clear, if one focuses on the names of the living that names aren’t definite descriptions. And it’s clear for the reasons that Mill gave. But Russell and Frege focused on a different class of names, and there it is.

1.9.2 Reason #2: Replacing co-referring proper names changes meaning

Let’s suppose that Mill is right—that names are mere labels and are not “connotative” (descriptive). In that case, there is some object—let O be that object—such that each of “Hesperus” and “Phosphorous” simply labels O. It follows that:

(H_1) “Hesperus is less massive than the Earth”

has the very same proposition for its literal meaning as

(P_1) “Phosphorous is less massive than the Earth,” namely:

(O_1) O is less massive than the Earth;

Indeed, it follows that, for any property psi,

(H^*_1) ‹Hesperus has psi›

has the very same proposition for its literal meaning as

(P_1) ‹Phosphorous has psi›,

namely:

(O^*_1) O has psi.

In particular, it follows that:

(IHP) “Hesperus is identical with Phosphorous”

has the very same meaning as

(IHH) “Hesperus is identical with Hesperus,”

namely:

(IO) O is identical with O.

But this all seems very wrong. IHP is non-trivial. H_1 and P_1 convey very different propositions (namely, O^*_1 and P^*_1); the same is true, for any property psi of IHP and IHH. Given these facts, it seems *de rigueur* to accept the Russell-Frege position that “Hesperus” and “Phosphorous” are synonymous with different definite descriptions (“the morning star” and “the evening star,” or whatnot).

But this reasoning is no good. We’ve already seen why. But let’s adapt what we said earlier to deal with the Hesperus-Phosphorous case. Let’s go back to the time when your friend was first telling you what “Hesperus” refers to. He points to a certain celestial body x, which is clearly the last one to disappear from the morning sky, and tells you that “Hesperus” refers to that thing—that is, that “Hesperus” refers to x. Thus, there is some x such that you know that x is the last celestial body to disappear from the morning sky and such that you are now being told that the semantic rule for “Hesperus” is:

(HSR) “Hesperus” refers to x.

Given these circumstances, an utterance:

(HSR_2) “Hesperus is hotter than the Earth”

would convey to you the information that:

(HSR_3) something is a last celestial body to disappear from the morning sky and any such thing is hotter than the Earth.

And HSR_2 would convey this even though its literal meaning would be the descriptively more impoverished proposition that:

(HSLM) O is hotter than the Earth

for some object O.

In general, for any psi, an utterance of (H^*_1) would convey the information that:

($HSRG_3$) something is a last celestial body to disappear from the morning sky and any such thing has psi.

And, echoing what we just said, an utterance of H^*_1 would convey that *even though* its literal meaning was simply:

(O^*_1) O has psi.

And for the same reasons *mutatis mutandis,* an utterance of P^*_1 would convey:

($PSRG_3$) something is a last celestial body to disappear from the morning sky and any such thing has psi

enough that the utterance's literal meaning would be O^*_1.

Russell and Frege overlooked the fact that semantics isn't learned in a vacuum and that, as a result, the non-semantic information through which semantic rules are learned is sometimes smuggled into the messages borne by utterances.

Reason #3: Replacing co-referring sometimes proper names changes truth-value

A person can know that:

(H_1) "Hesperus is less massive than the Earth"

expresses a truth without knowing that

(P_1) "Phosphorous is less massive than the Earth"

does so as well. Many a person who knows that

(IHH) “Hesperus is identical with Hesperus”

is true doesn’t know that

(IHP) “Hesperus is identical with Phosphorous”

is true as well.

This data is naturally accounted for by supposing that “Hesperus” is synonymous with “the morning star” and “Phosphorous” is synonymous with the evening star. And that’s just what Frege and Russell did suppose. But, as we’ve seen, this isn’t the right way to model the data and it isn’t the only way. The supposition that H_1 (IHH) and P_1 (IHP) coincide in literal meaning *does* model that data *provided* that the points concerning pre-semantic information made earlier are taken into account.

2.0 Kripke’s revival of Mill

In 1969, Saul Kripke gave three very clever lectures in which he definitively showed that the Russell-Frege analysis of proper names is wrong. Those lectures are now published under the title *Naming and Necessity*. He said many insightful things in the course of those lectures, many of which went far beyond the scope of semantics. He also said some things that are quite clearly false, but that, oddly, were nevertheless accepted and, indeed, turned into orthodoxies. Let’s start with the insightful, correct things that Kripke said.

2.1 The modal differences between names and definite descriptions[152]

For argument’s sake, suppose that:

(SOC) “Socrates” is synonymous with some definite description—with, let us suppose, “the great philosopher who died of hemlock poisoning.”

If SOC is right, then:

(1) “Socrates did not die of hemlock poisoning”

would be self-contradictory, since it would mean the same thing as

(2) “the great philosopher who died of hemlock poisoning did not die of hemlock poisoning.”

But (1) isn’t self-contradictory at all. We could imagine its being true. Suppose that historical research reveals that it wasn’t hemlock poisoning that killed Socrates—that the poison he drank was some other sort of poisoning. In that case, (1) would be false. This may not be likely, but there’s nothing incoherent in the idea of its happening; supposing this about Socrates isn’t like supposing that squares were really circles.

2.1.1 The modal differences between names and definite descriptions (continued)

A related point is that:

(3) “Socrates died of hemlock poisoning”

is not *analytic.* (An *analytic* statement is one that, given only what it means, *must* be true. So “squares have four sides” is analytic, but “Socrates was a wrestler” is not.) But it would be analytic if SOC were right, since it would then be synonymous with:

(4) “the great philosopher who died of hemlock poisoning died of hemlock poisoning.”

Why isn’t (3) a tautology? Because lots of people who know who “Socrates” refers to—who know that he’s the lead character in the *Republic,* or whatnot—may not know that he died of hemlock poisoning. Those people are not, for that reason, guilty of some semantic deficiency; they may know as well as anyone what the semantic rule governing “Socrates” is. There is some x (namely, Socrates) such that “Socrates” refers to x, and they know that. (4) is thus an empirical point.

2.1.2 The modal differences between names and definite descriptions (continued)

A consequence of these points is that:

(5) “it’s necessarily the case that the great philosopher who died of hemlock poisoning died of hemlock poisoning”

is true, whereas

(6) “it’s necessarily the case Socrates died of hemlock poisoning”

is false. Even if Socrates *did* die of hemlock poisoning, he might not have; he might fallen off a bridge on his way to the hemlock-drinking ritual.

For exactly similar reasons, “Benjamin Franklin” isn’t synonymous with “the inventor of bifocals.” For:

(7) it’s necessarily the case that the inventor of bifocals invented bifocals”

is true, but

(8) “it’s necessarily the case that Benjamin Franklin invented bifocals.”

Kripke’s basic point is that proper names don’t have the same *modal* properties as co-referring definite descriptions. The “modal status” of a sentence lies, not in whether it *is* true, but in whether it, or its negation, *can* be true. The proposition that squares have four sides is *necessarily* true. That is its modal status. To say that it is necessarily true is to say that it cannot be false. (1) and (2) express propositions that have different modal statuses and are therefore distinct. Thus, replacing the occurrence in (1) of “Socrates” with “the great philosopher who died of hemlock poisoning” results in a sentence that expresses a proposition different from that expressed by (1). Thus, “Socrates” isn’t synonymous with “the great philosopher who died of hemlock poisoning” or, by the same logic, with any other definite description. Given obvious extensions of this reasoning, no proper name is synonymous with any definite descriptions.

2.2 Kripke on identity and necessary a posteriori truth

To understand Kripke’s influential non-semantic claims, we must define four terms and also assimilate an important principle that Kripke discovered.

2.2.1 A priori vs. a posteriori knowledge

The first two terms are “a priori” and “a posteriori.” These expressions are adjectives, and they modify expressions denoting knowledge. A priori knowledge is knowledge that one either has, or can acquire, merely by virtue

of how one's mind is structured; it can be thought of as knowledge that is coded into one's cognitive structure or that follows from such knowledge. When you a buy a new computer, it typically comes "bundled" with new software. A priori knowledge can be compared to such software. A posteriori knowledge is knowledge that is not a priori.

2.2.2 Rigid vs. non-rigid designators

The remaining two terms are "rigid designator" and "non-rigid designator." For reasons that will emerge in due course, the distinction between rigid and non-rigid designators is meaningless unless definite descriptions are taken to be referring terms. So in this context, we'll suppose, just as Kripke does in *Naming and Necessity,* that definite descriptions are referring terms.

Consider a world W that is semantically just like ours except that, in it, Al Gore, not Barack Obama, is the U.S. President in 2009. In W, "the U.S. President in 2009" would refer to Al Gore. *So even though that expression would have the semantics that it has in our world,* it wouldn't pick out the same thing that it picks out in our world. In fact, given that, in W, Barack Obama isn't the president in 2009, if "the U.S. President in 2009" *did* pick out "Barack Obama" in W, it couldn't possibly have the same semantics that it has in our world.

But suppose that, in world W* that is distinct from both W and our world, "Barack Obama" referred to Joseph Biden. In that case, W* would *ipso facto* not be semantically identical with our world. In our world, there is some individual x such that x is Barack Obama and, as a matter of semantics, "Barack Obama" refers to x. And any language in which "Barack Obama" picked somebody else would to that extent semantically diverge from the specific version English actually spoken (even if it was otherwise just like it). Given that "Barack Obama" refers to Barack Obama, it couldn't possibly *fail* to refer to him in a world that is semantically like ours. It could fail to refer to him only in a world where expressions were not in all cases governed by the semantic rules that govern them in our world.

"Barack Obama" is a *rigid designator.* E "rigidly designates" x if there is no possible world that is semantically just like ours where E fails to designate x. Proper names are rigid designators. Definite descriptions are not. "The inventor of bifocals" refers to Bob Denver in some possible worlds, to Walt

Disney in others, and so on.

Sometimes *demonstrative* expressions (e.g., "that person") are described as "rigid designators." The reasons for this are explained in Chapters 6 and 9. Suffice it to say that they are not rigid designators. In fact, they are about as non-rigid as designators can be. "The inventor of bifocals" is non-rigid in the sense that there are possible universes in which it refers to people other than Benjamin Franklin. But "that person" refers to many different people in *our* world.

2.3 Identity and necessity

Suppose that x and y are contingently identical—that, in other words, they're identical but they don't have to be. By Leibniz's law, x and y have the very same properties. Obviously x is necessarily identical with x. Therefore, y has that property as well and, contrary to what we assumed, is necessarily identical with x.

In general, identities hold necessarily if they hold at all. It isn't a contingent fact that you are identical with you. If you exist, you can't fail to be you.

Some sentences seem to be counterexamples to this; for example, "the person standing over there is the person who owns the ice-cream store down the street." But that isn't really an identity claim. We know from Russell that its real meaning is: "something is both a person standing in that corner, unaccompanied by other people, and is also a sole proprietor of the ice-cream store down the street." Given any apparent exception to the principle that, when it holds, identity holds necessarily, reparsing that sentence shows it not to be an identity claim.

A sentence of the form "A is identical with B" expresses an identity claim *only* if "A" and "B" are rigid designators. Thus:

(IHP) "Hesperus is identical with Phosphorous"

is an identity claim. And so is

(WH) "water is H_2O"

and so is

(MT) “Mark Twain is Samuel Clemens.”

2.4 Why Kripke believes that there are necessary a posteriori truths

Kripke made it clear that “Hesperus” isn’t synonymous with “the morning star” or with any other non-rigid designator, and he showed the same thing *mutatis mutandis* to be true of “Phosphorous.” Given that Hesperus *is* Phosphorous, Hesperus can no more fail to be Phosphorous than it can fail to be Hesperus. So given that it holds, IHP holds necessarily.

Kripke says that “IHP” encodes an a posteriori truth. The idea is that it can’t be known a priori that IHP (or any other synonymous sentence) is true. It can be known only through astronomical observation. Given that IHP is necessarily true, it follows that IHP is both *necessary* and a posteriori.

2.5 Evaluating Kripke’s attempt to show that there are necessary a posteriori truths

This is not a good argument. We made it clear earlier in this chapter that, although it doesn’t have such a claim for its *literal meaning,* IHP *communicates* a proposition of the form *the phi is identical with the psi.* In other words, it (non-literally) communicates a proposition of the form: *something uniquely has phi, and anything that has phi also uniquely has psi.*

So IHP communicates a *contingent* truth. It communicates, as we saw, the proposition that is expressed by the sentence:

(MS) “the last celestial body to disappear from the morning sky is identical with the first to appear in the evening sky” (which we’ve abbreviated as: “the morning star is identical with the evening star”).

The proposition meant by MS is:

(MSP) something is uniquely a last body to disappear from the morning sky, and anything having that property also uniquely has the property of being a first thing to appear in the evening sky.

MSP is a posteriori. But it’s also contingent. It wouldn’t be true if matter were differently distributed.

Kripke is quite right that IHP has a necessarily true proposition for its meaning. But that proposition is also analytic and knowable a priori.

Here's why. We've agreed with Kripke that, for any property phi, "Hesperus" isn't synonymous with ‹the phi.› And we've also agreed with Kripke, and with commonsense, that "Hesperus" is a referring expression—that, in other words, there is some object x such that "Hesperus" picks x out. Trivially, either "Hesperus" picks out this object through the mediation of a description or it doesn't. In other words, supposing that O is what "Hesperus" picks out, it either is, or is not, by virtue of O's fitting some description that "Hesperus" picks out it. If it is, then "Hesperus" is *ipso facto* a non-rigid designator—a definite description. For supposing that the description that O must satisfy is *unique thing having phi,* "Hesperus" must be synonymous with ‹the phi.› But we know, from Kripke himself, that "Hesperus" is not synonymous with any such expression. So "Hesperus" does *not* refer to O through the mediation of a description. It picks O out—but *not* because it has this or that property. It picks it out because the semantic rule for "Hesperus" is to the effect that, no matter what properties O has, "Hesperus" is to pick it out—because, in other words, "Hesperus" picks out O, this not being contingent on O's satisfying any condition (other than its being O). Thus, there is some object x (namely, O) such that the semantic rule for "Hesperus" is to the effect that, no matter how O is, "Hesperus" picks it out. In other words, "Hesperus" *picks out O directly.* It is *directly referential.*

A consequence is, for some object x (such that x is Hesperus) such that, in virtue of having the form, ‹Hesperus has psi›, a sentence is true if and only if x has phi. For exactly similar reasons, "Phosphorous" is directly referential. It co-refers with "Hesperus." Thus, there is some object x (such that x is Hesperus and Phosphorous) such that, in virtue of having the form, ‹Hesperus has psi›, a sentence is true if and only if x has phi and such that, in virtue of having the form, ‹Phosphorous has psi›, a sentence is true if and only if x has phi. Thus, ‹Hesperus has psi›and ‹Phosphorous has psi› *have the very same literal meaning.* In general, inter-substituting "Hesperus" and "Phosphorous" doesn't change literal meaning. (This is entirely consistent with the fact such inter-substitutions may dramatically change suggested meaning, the reason being that one may learn what those two words refer to under very different circumstances and thus in connection with very different bodies of

descriptive data. The way the "referent is fixed" will obviously affect the connotation of sentences containing the word in question, even though it won't affect the literal meanings of such sentences.) Given that such inter-substitutions preserve meaning, it follows that:

(IHP) "Hesperus is identical with Phosphorous"

has the same literal meaning as

(IHH) "Hesperus is identical with Hesperus"

(even though, given how it is typically made known what these words refer to, they have very different non-literal connotations). We've agreed that, for some x, each of ‹Hesperus has psi› and ‹Phosphorous has psi› is true just in case x has phi. It follows that, for some x, each of IHP and IHH is true just in case x has the property of being identical with x. Thus, for some x, IHP means:

(XSI) x is x.

And XSI is an analytic, a priori knowable proposition if ever there was one.

There is no a posteriori necessary truth. It appears otherwise only if one fails to distinguish sentences from the propositions that are their meanings. The gulf between the two can be large. This is but another example of the fact that metaphysics can't be read off of sentential structure. Kripke's disregard for this shibboleth of analytic philosophy comes as a shock. But that shock is nothing compared to the one we're in for.

3.0 A cure for Nathan Salmonella

Kripke (1979) wrote an article in which he put forth a new argument (one not stated in *Naming and Necessity*) as to why the Russell-Frege analysis of the non-triviality of IHP fails. Pierre is a monolingual speaker of French. He reads many books about London (which he knows as "Londres"). On the basis of the information contained in those books, he draws the rational inference that London is a beautiful city. Expressing this belief, he says:

(1) "Londres est jolie."

It is to be emphasized that, given the data at his disposal, it is *rational* of

Pierre to accept (1).

Our story continues. Pierre is kidnapped and forced to live in a terrible part of London. He doesn't know that he's living in the city known to him as "Londres." He learns English, but he learns it as one learns a first language. In other words, he doesn't learn what English words mean by learning what their French translations are—he learns them the way that a native speaker of English learns them. He learns what "snow" means, not by learning that it co-refers with "neige," but by learning a certain crystalline substance that falls on rooftops in the winter is called "snow." And he learns what "London" refers to in the same way. Given the data at his disposal, it is *rational* of Pierre to believe that the sentence:

(2) "London is not pretty"

expresses a truth—a truth to the effect that the city he is living in is *not* pretty. And Pierre does assent to it, and he utters that very sentence, knowing what it means.

Of course, (2) is the (English translation of the) ***negation*** of (1).

"London" is the translation of "Londres." So if "London" is synonymous with some definite description—with, let us suppose, "the most populous city on the Thames"—then "Londres" is synonymous with the French translation of that definite description. But if *that* were the case, then (1) would mean the same thing as:

(1_R) "the most populous city on the Thames is pretty"

and (2) would be mean the same thing as:

(2_R) "the most populous city on the Thames is *not* pretty."

Pierre clearly *isn't* in the same category as somebody who accepts *both* (1_R) and (2_R). So even if, as Russell and Frege thought, "London" was synonymous with some definite description, that wouldn't explain the fact that (1) and (2) communicate such different propositions. In other words, it would fail to account for the *cognitive values* of those sentences. (The "cognitive value" of a sentence is what it communicates.) Since the whole

point of the Frege-Russell theory that proper names are (abbreviated) definite descriptions is to account for the cognitive values of sentences containing proper names, that theory fails to do what it's supposed to do.

3.1 A cure for Salmonella (continued)

But Kripke's parable exposes a problem much more profound than anything having to do with the semantics of a proper name. It's a datum that Pierre understands each of (1) and (2), that he assents to each, and sincerely utters each. So it's a datum, or follows from data, that Pierre is *rationally* accepting some sentence *and* its negation.

This seems like a problem. A paradox. What's the solution? Pierre is rational. No denying it. And Pierre understands each of (1) and (2). No denying that either. (To be sure, he doesn't know as much *about* the thing that is referred to as "London"/"Londres" as some other people. But so what? If I know who it is that "Larry" refers to, then I know the *semantics* of "Larry." Obviously I don't know who Larry is as a person as well as his therapist. But I know the *semantics* of "Larry" every bit as well.)

Kripke says that he doesn't know how to solve the problem. He says only that he's inclined to suspect that "the apparatus of propositions breaks down here." He doesn't say what he means by this. Nor does it matter, since, whatever it means, it's false, as we'll see.

Nathan Salmon says that, given Kripke's parable, it follows that, at a given time, one *can* rationally accept a proposition *and* its negation.[153] Pierre would be rational, given what we've said so far, to believe that "Londres ≠ London" is a true sentence. Thus, there is some x (namely, London) such that, according to Salmon, Pierre can rationally believe that: x ≠ x—that is, that x is not self-identical. Thus, Salmon's position is that one can rationally believe that London is not self-identical—that, in fact, any given object is not identical with itself.

3.2 Why, contrary to what Nathan Salmon ways, it is not rational to believe both P and not-P

Salmon's solution is not only desperate, but utterly incoherent. What is

rationality? It's the ability to draw the consequences of the information at one's disposal. No truth is a consequence of its own negation, and every truth is incompatible with its own negation. Given that *P and not-P* is always false, it can never be rational to accept it or, therefore, to infer it. Salmon's position is inconsistent with the very meaning of the word "rational."

Salmon sometimes defends his position by saying that one can rationally accept *P and not-P* only if one doesn't know that what one is accepting is a proposition *and* its negation. But this makes no sense. You are irrational if you believe *P and not-P* or if you grasp both P and not-P and yet fail to generate the additional belief that they're inconsistent with each other. For acceptance of *P and not-P* to be rational, it would be necessary that the data at your disposal not make it clear that you were accepting some proposition and its negation, which would be possible only if you didn't know what it was you were accepting. And this seems to be what Salmon holds.

But, contrary to what Salmon evidently thinks, you can't accept a proposition without knowing what it is. The counterexamples to this are spurious. Suppose that Smith believes anything that he believes Jones to believe. If Jones believes that Lima is the capital of Peru, does it follow that Smith also believes it? No. What follows it is *if* Smith were to know that Jones believed it, *then* Jones would believe it. In order to accept a proposition, it is not enough that it be *described* to you. You must actually grasp it. Another story will make this clear.

Person X is never convinced of anything unless it actually is true. This has been verified, and I know this. Smith and I have a mutual friend, Green, who has been missing for the last three weeks. One day Smith tells me that he knows where our friend Green has been for the last three weeks. I have no idea where Green has been. I couldn't even guess. But Smith doesn't tell me where Green has been—only that he knows where he's been. So there is some proposition P such that, if correct, P says where Green has been and such that Smith accepts P. I know that whatever it is that Smith believes about Green's whereabouts is correct. But do I, under these circumstances, accept the proposition that identifies his whereabouts? No. If I were told what that proposition was, I *would* accept it. But I do not now do so. I don't know what the right value of P is.

Suppose that, as Smith knows, Green has been in Spain for the last three weeks but that I'm not told this. Under these circumstances I don't accept the proposition: Green has been in Spain for the last three weeks. I *would* accept it if I knew that Green accepted it. But I don't yet know this, and I don't yet accept it. The proposition that I *do* accept is *this: I know that, whatever Smith believes to be the proposition describing Green's whereabouts over the last three weeks, that proposition must be correct.* But *that* proposition is very different from *Green has been in Spain for the last three weeks.* And I don't yet accept the latter.

What this shows is that one cannot accept a proposition without knowing that one is doing so. When it *seems* as though one does this, one accepts some distinct proposition that describes some unknown proposition. If you accept a proposition, you know it. So if, for some P, you accept both P and not-P, you know it. But if you know it, you're not rational, since anyone who makes a faulty inference is *ipso facto* irrational (to that extent), and any inference whose conclusion is *P and not-P* is *ipso facto* faulty.

3.2.1 Why, contrary to what Nathan Salmon says, it is not rational to believe both P and not-P (continued)

Salmon was driven to accept this preposterous view by his failure to take into account even the most basic points concerning the relationship between thought and language. Pierre doesn't learn what "London" refers to in a vacuum. He is air-dropped into a horrible part of a certain city. There is garbage all around. His captors are rude to him, and so are the people with whom his captors associate. (The only Londoners with whom Pierre's captors consort are, uncharacteristically for Londoners, not very nice people.) He is then told: "Welcome to London! This is your new home." Let us suppose, in fact, that one of his captors is a philosopher of language who, like all those in that field, has a penchant for making things explicit. So, when first introducing Pierre to his new home, he says: "there is a certain city x, such that x is the city you are now occupying— this being the city whose dirty, garbage-ridden streets you are now beholding—and such that 'London' refers to x."

What Pierre is being told, in being told this, is something along the lines of:

(L_1) there is some vile, squalid, garbage-infested city x such that I am now in x and such that I was dragged here against my will and such that, because of x's obvious horridness, I am now depressed; moreover, "London" refers to x.

An inevitable consequence of the fact that these are the circumstances under which Pierre learns what "London" refers to is that what an utterance of

(2) "London is not pretty"

would impart (not semantically encode) is something along the lines of:

(L_2) there is some vile, squalid, garbage-infested city x such that I am now in x [etc.] and such that "London" refers to x; moreover, x is not pretty

and, in general, that for any property psi

(L_n) ‹London has psi›

would impart (not semantically encode) something along the lines of:

(L^*P_2) there is some vile, squalid, garbage-infested city x such that I am now in x [etc.] and such that "London" refers to x; moreover, x has psi.

We've already made it clear why, even though L_n would *convey* L^*P_2, Pierre wouldn't be under any illusions as to what the former literally means. It's also obvious that, since it's through L_1 that Pierre knows what "London" refers to, it is rational of him to accept (2).

Let's think back to the time when Pierre, not yet having been abducted, was still in Paris. As we've said, Pierre has a lot of information about the city he knows as "Londres" and he draws the only reasonable inference that, given that information, he can draw; namely, that it is a pretty city. For he has read in various books that he knows to be authoritative that:

(##) There is a certain city x such that x is more populous than any other in Britain, such that is the Capital of Britain, such that there are beautiful buildings in x, such that x's streets are generally very clean, there isn't much violent crime in x, there are beautiful buildings in x, etc.; moreover, "Londres" refers to x.

Given that it is through some proposition along the lines of ## that Pierre

knows what “Londres” refers to, it is inevitable that an utterance of:

(1) “Londres est jolie”

would convey to him something along the lines of:

(##^) There is a certain city x such that x is more populous than any other in Britain [etc.] and “Londres” refers to x; moreover, x is pretty.

And given that (1) communicates (##^) to Pierre, it’s clear, given the information at his disposal, that he’s rational to accept (1).

Here’s what’s going on. Pierre knows what “London” refers to and he also knows what “Londres” refer to. He doesn’t know that they refer to the same thing, since the body of descriptive information that he uses to attach “London” to the right thing is different from the body of descriptive information that he uses to attach “Londres” to that thing. This is consistent with the supposition that he knows what “Londres” and “London” refer to. (I may not know that my friend Smith is also the masked bandit and, therefore, that “the masked bandit” refers to Smith. But I know who “Smith” refers to—he refers to my amicable next-door neighbor. And I also know who “the masked bandit” refers to—it refers to the cape-wearing buffoon frequently featured on the evening news.) Given that Pierre speaks both French and English, he understands (1) and (2). He rationally assents to both, even though the one is the negation of the other. This is *not* because he ratio-nally accepts some proposition *and* its negation. *Nor* is it because “the apparatus of propositions breaks down,” whatever that means. No—it’s because, given the information *through* which he learned the relevant semantic rules (the ones that, for some x (namely London), are to the effect that “Londres” refers to x and that “London” refers to x), (1) and (2) communicate very different propositions to Pierre. Those propositions don’t entail each other, and they are, in addition, supported by very different bodies of data.

There is no proposition that Pierre both accepts and rejects. Salmon is just plain wrong to think otherwise. And his contention that one *can* rationally accept propositions and their negations involves a number of rather obvious fallacies. First, sentences are not propositions. Suppose that sentence S_1 encodes proposition P_1 and that sentence S_2 encodes the negation of P_1. Given *only* that information, it doesn’t follow that a person who understands

S_1 and S_2 can't rationally accept both of those sentences. For it could be that, given the information through which that person learned the relevant semantic rules, the *propositions* conveyed to him by those sentences *could* both be true. Salmon agrees that, given only that the proposition encoded in the one sentence is the negation of the proposition encoded in the other, one can rationally accept both sentences. But *he* thinks this is because one can rationally accept *propositions* that are inconsistent with each other; Salmon thinks that one accept the *propositions* P *and* not-P *without being irrational.* This is a deeply irrational belief. Irrationality is acceptance of incoherencies, such as P and not-P.

Salmon holds, correctly, that Pierre is rational in accepting both (1) and (2). According to Salmon, it follows from this that Pierre is rationally accepting some proposition *and* its negation. But that would follow *only* if what (1) and (2) conveyed to Pierre were confined to what they literally mean. But what sentences convey isn't always confined to what they literally mean. It's seldom confined to it. (See Chapters 9 and 18.) Every semanticist knows this. It's unclear why Salmon doesn't know it or why, if he does, he didn't see its bearing on this issue.

3.3 Some final remarks on Frege's conception of reference: Frege's distinction between sense and reference are a distortion of the distinction between pre-semantics and semantics

We've seen that if a proper name N refers to some object x, then, in virtue of having the form ‹N has psi›, a sentence S is true just in case:

(i) x has psi.

It immediately follows, as we've noted, that Frege's analysis of proper names is false. According to Frege, proper names, and indeed all referring expressions, have "senses." If a referring term M refers to y, M's sense is some property phi that y uniquely instantiates. And, if Frege is right, in virtue of having the form ‹M has psi› a sentence S* is true just in case:

(ii) phi is uniquely instantiated; moreover, any such instance has psi.

No sooner is Frege's analysis made clear than its incoherence is plain to see. (ii) is an existential generalization, and the *reason* for this is that it contains what Frege regards as a referring term. Thus, if Frege's analysis of proper names is right, they are semantically indistinguishable from existential quantifiers and, therefore, *are* such quantifiers. But Frege says they're referring terms. Since, therefore, Frege's position is, in effect, that proper names both are and are not quantifiers, it is incoherent.

3.3.1 Indirect reference = quantification = non-reference, direct reference = only true form of reference

Frege's position is that names refer *indirectly.* In other words, they refer through the mediation of a sense, and that sense is what makes it into the propositions meant by sentences containing those names. The referents of those names don't make it into those propositions. By virtue of containing proper names of those objects, sentences containing those names merely *describe* their referents. But we've just seen that "indirect reference" is really just quantification.

Frege's distinction between "sense" and "reference" is an incoherent way of describing the distinction between semantics and *pre*-semantics. There is some entity O, namely London, such that the proposition literally meant by 2 is true just in case:

(ONP) O is not pretty,

but such that, because of how Pierre *learned* what "London" refers to, what 2 *conveys* to him is (L^*P_2).

Frege would have said that *the* sense of "London" is some property uniquely had by London (e.g., the property of being a city that is vile, squalid, etc.). But that isn't true. One's knowing what "London" refers to consists in there being *a* property phi—it doesn't matter which one—such that one knows that phi is uniquely instantiated and that "London" refers to any instance of it. There is no *one* property phi such that one's knowing of what "London" refers to consists in one's knowing that phi specifically is uniquely instantiated and that "London" refers to an instance of it.

And yet Frege constantly speaks of *the* sense of "Socrates," "London,"

etc. Although this view is patently false, Frege has a good reason for holding it. He thinks that if "London" had as many senses as there were people who knew what it referred to, then "London" would be useless as a device of interpersonal communication. When I said "London is pretty," I'd mean one thing (e.g., the most populous city in the UK is pretty) and what *you'd* mean by that sentence would be something else (e.g., *the city where grandma lives is pretty*).

But we've already seen why this is spurious. There is some x such that I know that x is the most populous city in England and such that I know that "London" refers to x and, finally, such that, when I say "London is pretty," what I am affirming, and what other English-speaking auditors (who know that I speak English) know me to be affirming, is:

(LP) x is pretty.

And there is some x such that you know that x is where your grandmother lives and such that "London" refers to x and, finally, such that, when you say "London is pretty," what you are affirming, and what other English-speaking auditors (who know that you speak English) know you to be affirming is

(LP) x is pretty.

So while it's a fact that "London" has one sense for me and a different sense for you, that doesn't prevent us from using it to make ourselves understood to each other. And the reason is that these senses don't make it into what we are affirming or what we believe ourselves to be affirming.

Bearing these points in mind, let's wrap up our discussion of Pierre. Pierre has a thorough grasp of the semantics of "London." For him, the sense of "London" is a *squalid and horrible city where my kidnappers have forced me to live,* or some such. There are others who grasp the semantics of "London" but don't associate that property with that expression. The sense of "London" varies from person to person. (We've just explained why this does not undermine people's ability to use that expression to make themselves understood to others.) This is because what the sense of "London" is for a given person depends, not on its literal meaning, but on the information through which that person *grasps* that literal meaning. Sense, therefore, is *pre*-semantics, not semantics.

3.4 Salmon on non-existence and mythical objects

Nathan Salmon holds that one *can* say of a given thing x that x exists or that it does not exist. Indeed, Salmon says that one can *correctly* say of a given thing x that x does not exist. Of course, for one to *correctly* say of a given thing x that it doesn't exist there must exist some thing such that one is correctly saying of it that it doesn't exist. So there must exist something that doesn't exist. Since this isn't possible, Salmon is wrong.

But arguments like the one just given have been around since Russell first produced them over a hundred years ago; and Salmon, being a devotee of Russell's works, is well aware of these arguments. So why does he hold that one can correctly attribute non-existence and existence to objects?

Salmon has a couple of reasons. The sentence:

(1) "Socrates is Socrates"

doesn't mean the same thing as

(2) "Socrates is the main character in most of Plato's dialogues."

Thus, "Socrates" isn't synonymous with "the main character in most of Plato's dialogues" or, by obvious generalizations of this argument, with any other ***definite description.*** This argument, duly generalized, shows that proper names are not in general synonymous with definite descriptions. From this it follows that proper names are mere labels. "Socrates" no more *describes* "Socrates" than your social security number describes you.[154]

Definite descriptions describe their referents. To be referred to by "the tallest person in the room," you have to have certain characteristics. You must uniquely be a tallest person in the room. If a taller person walks in the room, that expression won't pick you out. But, whatever your name is, there isn't any way you have to be for that expression to pick you. It doesn't matter how tall or short you are; what your job is; etc. Thus, names are mere labels.

It follows that, if you say "Bill Jeffries is tall," there is some individual x such that "Bill Jeffries" refers to x and such that for your statement to be true, it is necessary and sufficient that x be tall—it being entirely irrelevant what other properties x has. The same thing *mutatis mutandis* obviously holds with respect to "Socrates," "Winston Churchill," etc.

Thus, if N is a proper name (that isn't a make-believe name or is otherwise one that doesn't pick anything out) and psi is any property, there exists some object x such that ‹N has psi› is true iff x has psi—it being irrelevant what other properties x has or lacks.

But now we have a problem. "Winston Churchill" isn't an "empty" (non-referring) name. It isn't in the same category as "Rooster Cogburn." And "Winston Churchill was bald" is meaningful. But Churchill no longer exists. Thus, there does not exist an object x such that "Churchill" refers to x.[155]

Some people deal with this by trying to revive the idea that proper names are definite descriptions. But this isn't a good idea, since the arguments just given are clearly probative. Salmon rightly avoids this path. Others deal with it by saying that, since existence is "four-dimensional," nothing ever ceases to exist. But Salmon rightly rejects this tortured and implausible view.

Unfortunately, the way Salmon deals with the problem is no more plausible than the two views that he rightly rejects. Salmon says that it *is* meaningful to say of a given thing (e.g., Churchill) that it doesn't exist. Salmon's position is that "Winston Churchill" *does* have a referent—a non-existent one.

If Salmon were right to say that one could say of specific non-existent things that they were non-existent, it would follow that one could say of specific non-existent things that they were existent. If a proposition can be affirmed, so can its negation. (Of course, one couldn't *correctly* say of a given non-existent thing that it existed. But one could still say it—just as one can say, albeit falsely, that Barack Obama is eight-feet tall.) For similar reasons, it would follow that one could meaningfully (though, of course, falsely) say of a given existing thing that it didn't exist and also that it did exist.

Of course, in saying that "Winston Churchill" has a non-existent referent, Salmon is saying that there exists something x such that "Winston Churchill" refers to x and such that x doesn't exist. And, of course, that statement is self-contradictory—so far as, given the points made earlier, it isn't completely meaningless. For some reason, Salmon isn't bothered by the consequence of this theory.

Salmon is quite right to hold that proper names are mere labels. But,

contrary to what he thinks, that fact doesn't license his desperate and incoherent views about existence. Consider the sentence:

(S_1) "JMK plays tennis."

This sentence is meaningful. In other words, there is some ***proposition*** that it expresses. Let P_1 be this proposition. We've seen that there is some x such that "JMK" refers to x such that S_1 is true just in case x plays tennis—that, in other words, P_1 is true iff x plays tennis. Salmon knows this, and he concludes from this that JMK is a veritable *constituent* of P_1. In other words, there is some proposition that I am actually a part of. This is a very contrived view, given that I'm a spatiotemporal entity, whereas propositions are not.

In any case, it isn't the right view. Under what circumstances is P_1 true? First of all, the property of being a tennis player must be instantiated. (If nobody plays tennis, P_1 is false.) For obvious reasons, it's also necessary that the property of being identical with JMK be instantiated. Let PTP and PJMK be the property of being a tennis player and the property of being identical with JMK, respectively.[156] Finally, let PJMT be the property of being identical with JMK *and* being a tennis player.

Supposing that P# is the set containing PTP, PJMK, and PJMT, and not containing anything else, the members of P# are collectively instantiated if and only if P_1 is true. There is thus no reason not to say that P_1 *is* P# and that for P_1 to be *true* is for all of P#'s members to be instantiated.

The *property* of being identical with JMK will exist even after I no longer do, and it existed before I was born. It wasn't until I was born that PJMK was *instantiated;* and when I die, it will cease to be instantiated. But that's irrelevant: properties pre-exist and post-exist their instances.[157]

This analysis is consistent with the view (held by Salmon and myself) that "JMK" refers to me, not by *describing* me, but just by labeling me. Supposing, as we are, that P# is identical with P_1, S_1 has exactly the truth-conditions that Salmon believes it to have. There is some x (namely, JMK) such that for P_1, as we've analyzed it, to be true, it is necessary and sufficient

that:

(TCS_1[158]) x likes tennis

After I die, there won't exist an object of the kind just described. But that's irrelevant. After I die, P_1 won't be true; it will be false to say, at that point, that JMK plays tennis, the reason being that, sadly, he won't be doing anything. But our analysis is consistent with that. After I die, PJMK won't be instantiated.

It *will* be true at that point that JMK *did* play tennis. But our analysis is consistent with that fact as well.

The sentence:

(S_2) "JMK used to play tennis"

says the same thing as:

(S_3) the proposition *that JMK plays tennis* used to be true.

According to our analysis, S_3 is true iff PTP, PJMK, and PJMT *used* to be collectively instantiated. And if uttered after my passing, S_3 will be true exactly if, at the time of that utterance, those three *used* to be instantiated. Exactly similar points hold of the propositions expressed by "Churchill was bald," "Socrates couldn't read," etc. Given each of these propositions, Salmon says that it contains a non-existent object. Salmon also says that each of these sentences has a "non-existent proposition" for its meaning. He grants that "Churchill was bald" is meaningful and, therefore, that it encodes a proposition. But he says that this proposition doesn't exist. So that sentence is meaningful by virtue of encoding a non-existent proposition.

This is disastrous—and unnecessary. Unlike Salmon's analysis, ours doesn't require that there exist anything that doesn't exist. Ours doesn't require *any* violations of logic or commonsense. Our analysis makes only three demands: (i) that there is such a thing as being me; (ii) that to be such a thing as being a tennis player; and (iii) that there be such a thing as me playing tennis. Each of those demands is met.

I happen to be a Platonist, and I happen to think that each of those three things is a non-spatiotemporal object. But that's irrelevant. This analysis goes through as long as those three things exist. It doesn't matter how they're thought of.

Salmon doesn't even consider anything like this analysis. Instead, he holds that, since proper names are labels, it follows more or less directly that there exist things that don't exist. But surely it's a very bad idea to throw out the most basic principles of logic and to overhaul our most basic beliefs about reality, just so as to validate some semantic result concerning a certain class of nouns. Moreover, Salmon's way of validating that fact isn't even the right one, since, as we've seen, that fact is easy to validate within a framework that is entirely congenial to commonsense and its philosophical derivatives.

Salmon has another, similar reason for holding that one can meaningfully say of a given non-existent object that it exists (or that it doesn't) and also that one can meaningfully say of a given existent object that it doesn't exist (or that it does). It's obvious that:

(ZA) "Zeus is very angry right now"

isn't in the same category as "blarg blurbo gloxo." Utterances of ZA communicate propositions. We don't *believe* the proposition communicated by such utterances. But all that is relevant here is such that utterances do in fact convey propositions.

Note: ZA is a sentence of English. But it will help if, whenever necessary, you let ZA represent its Ancient Greek translation.

There was a time when people actually believed ZA (or its Ancient Greek equivalent rather). There may be people who believe it to this day. In general, there are proper names that don't refer to anything. Here I'm not referring to names, such as "Socrates," whose referents died long ago. I'm referring to genuinely "empty" names.

Supposing that N is an empty name, one would think that:

(NS) ‹N is a fierce warrior›

would fail to convey anything—that, so far as what it communicates is concerned, it would be in the same category as:

(ES) "__ is a fierce warrior."

But this isn't the case. A story will make this clear. Your friend Smith, who is generally pretty credible, tells you a story about a friend of his who has the unusual name of "Gigantus." You don't know Gigantus. When Smith first tells you about him, he says:

> (SG[159]) "My friend Gigantus is truly extraordinary. Gigantus was the first person to translate the complete works of Shakespeare into Albanian. Gigantus used to be a professor, but he now lives in the woods. He actually lives in a hollowed out tree trunk. For a while he was the best paid male model in Europe [etc.]."

Smith entertains you with stories about Gigantus for several weeks. Of course, these stories are replete with statements of the form ‹...Gigantus...› (e.g., "Gigantus recently read *War and Peace,*" "You'd really like Gigantus if you met him," etc.). It goes without saying that, at least from your viewpoint, these statements aren't like ES, let alone "blarg blurbo gloxo." They're quite as meaningful as "Socrates was a fine orator."

Then one day Smith tells you that there is no Gigantus. Smith made him up. It was all a hoax. There is no individual x such that "Gigantus" refers to x. It follows that:

(GW) "Gigantus recently read *War and Peace*"

fails to have a proposition for its literal meaning. So far as that sentence (or pseudo-sentence) has a meaning, its meaning is comparable to ES's. (It's meaning, it would seem is a propositional function, viz. __ *recently read War and Peace.*) So how are we to explain the fact that, as far as what it communicates is concerned, GW is in the same category as "Socrates was tall," as opposed to "blarg blurbo gloxo"?

Here is my answer. When you first encountered it, the name "Gigantus" was embedded in SG. (Remember that SG was told to you by your friend Smith.) It immediately follows that, when you first heard GW, what that utterance *communicated* to you was some message quite a bit richer than its (as it turns out, nonexistent) literal one. What that utterance of GW communicated to you was along the lines of:

(GWC) There exists a certain extraordinary individual x such that "Gigantus"

is x's name, such that at this moment my friend Smith is telling me about x, such that x lives in the woods in a hollowed tree-trunk, even though x used to be a professor [etc.]; moreover, x recently read *War and Peace.*

GWC is not GW's *literal meaning.* This cannot be emphasized enough. But it is what, because of your circumstances, GWC conveys to you.

The point is that, even though GW doesn't have *any* proposition for its literal meaning, there is no difficulty accounting for the fact that, empty of literal meaning though they be, utterances of it are replete with non-literal communicated meaning. We've already seen why much of what sentence-utterances communicate isn't literal meaning—why much of it is really *pre*-semantic content that reflects the idiosyncrasies of our knowledge of semantics and that we have some tendency to mistake for semantic content.

Nowhere in all this did we suppose that "Gigantus" is synonymous with a definite description. We didn't have to make that supposition to account for the data. Nor would it have been advisable for us to do so. "Gigantus" isn't a definite description. It's a proper name, albeit an empty one. Whatever it is that utterances of ‹...Gigantus...› might communicate, it isn't because "Gigantus" is a definite description that they communicate it.

Notice that our analysis is a very conservative one. We didn't posit any new entities. We didn't assume facts not in evidence (e.g., we didn't presuppose the truth of any uncorroborated hypotheses or otherwise prejudge open questions).

Now let's look at what Nathan Salmon says about empty proper names, like "Gigantus." He says that socalled "empty" proper names *aren't* empty. He says that "Zeus" *does* refer to a Greek God. He says that "Vulcan" *does* refer to an actual planet. ("Vulcan" was the name given by astronomers in the 19th century to a planet that they posited, but that turned out not to exist.) Salmon concedes that Zeus and Vulcan don't exist in space-time. If you flew around the solar system, you'd never come across Vulcan; if you climbed Mt. Olympus, you'd never see Zeus. But they do exist—they exist, Salmon says, as abstract entities.

Salmon's primary reason for accepting this view is that, since he thinks that names are mere labels, he (rightly) doesn't see it as an option to see "Vulcan" and "Zeus" as being synonymous or identical with definite

descriptions. But he finds it puzzling that, for any psi, utterances of ‹Zeus has psi› and ‹Vulcan has psi› are so rich in communicated meaning. He doesn't consider the proposal that we put forth a moment ago, and instead tries to solve the problem by saying that, in fact, "Vulcan" and "Zeus" *do* refer to things.

This isn't necessary. It's also pointless. What is communicated by sentence-utterances of the form ‹Joseph Biden has psi› and ‹Charles Keating has psi›, and other sentence-utterances containing no non-empty names, diverges dramatically from what such utterances literally mean. This is a well-known fact. So what it is that "Joseph Biden is a democrat" conveys cannot be explained solely in terms of what it literally means. So even if Salmon were right to say that there is some actual God to whom "Zeus" refers, that wouldn't explain any of the relevant data—it wouldn't explain why utterances of "Zeus has a beard" communicate what they do.

There's also the fact that Zeus, if he exists, isn't an abstract object. If he's anything, he's a God. If you pay me $1,000,000 to find Zeus, and all I find is an abstract object, I should give you your money back.

Salmon says that "Vulcan" is a planet—albeit a mythical one. Planets are giant hunks of rock in outer space. "Mythical" planets aren't hunks of rock. So they aren't planets. The expression "mythical planet" is in the same category as "Mr. Right." Its semantic function isn't to refer to anything and is instead to help us abbreviate otherwise long and unwieldy statements.

Salmon published an article, titled "Mythical Objects,"[160] in which he attempts to corroborate these views of his. Here is the argument presented therein:

> (MO[161]) Imagine the following. Smith and Jones both believe that Zeus exists, and they both think that he lives on Mt. Olympus and can hurl lightning bolts. One day Smith wakes up with a terrible headache. He thinks that Zeus gave it to him. That same day, Jones can't find his car. He thinks that Zeus destroyed it. Smith and Jones tell each other about their misfortunes, and they're in agreement that Zeus is indeed the culprit in both cases. It is thus a fact that:
>
> (1#) Smith and Jones believe some one deity (who lives on Mt. Olympus

and can hurl lightning bolts) to be responsible for Smith's headache and also for the disappearance of Jones' car.

(1#) is ambiguous, as it could mean either:

(2#) there exists some deity x (such that x lives on Mt. Olympus and can hurl lightning bolts) such that Smith and Jones both believe that x destroyed Jones' car and also gave Smith a headache

or

(3#) Each of Smith and Jones believes there to be some deity x (such that x lives on Mt. Olympus and can hurl lightning bolts) such that x is identical with Zeus, such that x destroyed Jones' car, and such that x gave Smith a headache.

(3#) cannot be the right disambiguation. (3#) says that Smith believes some one deity to be responsible for those two misfortunes, and that Jones believes some one deity to be responsible for them. But (3#) leaves it open whether the deity that Smith has in mind is identical with the one Jones has in mind. That means that (2#) is the correct disambiguation, and therefore, that there exists some deity who lives on Mt. Olympus, etc.

This argument evaluated: MO presupposes the truth of its own conclusion. Smith's belief that Zeus exists is identical with his acceptance of something along the lines of:

(Z) there exists a deity who lives on Mt. Olympus and hurls lighting bolts (etc.),

the same being true of Jones' corresponding belief.

Thus, Smith's belief that Zeus is responsible for both misfortunes is identical with his acceptance of something along the lines of:

(ZMF) there exists a deity x who lives on Mt. Olympus and hurls lighting bolts (etc.); moreover, x is responsible for both misfortunes

the same being true of Jones' corresponding belief.

And the statement that Smith and Jones "believe some one deity [etc.] to be responsible for Smith's headache and the disappearance of Jones' car" is just a loose way of saying that Jones and Smith both accept ZMF.

If there is no Zeus, there is no one entity that Smith and Jones both blame for these misfortunes. Each individual may *think* that he and the other are both blaming some one deity. But they're wrong. And (1#) is either false or it's a misleading way of saying that Jones accepts ZMF and so does Smith. Salmon's argument therefore fails *unless* it's assumed that Zeus exists. But if this is assumed, the argument assumes the truth of its conclusion and is therefore broken.

4.0 Natural kinds and the modality of identity-claims

Consider the statement:

(WH) "water is H_2O."

According to Kripke, WH encodes a necessary a posteriori truth. Here is his argument.

> (KA[162]) It obviously can't be known a priori that water is H_2O. Look at a glass of water. Can it be known through "pure reason" or through sheer logic chopping that it consists of H_2O molecules? No! Thus, WH is a posteriori.
>
> We've discovered that water is H_2O. Given this, suppose that we discovered some substance that was macroscopically just like water, but upon examination turned out not to be H_2O. Would we say that it was water? Would we say that water *can* be something other than H_2O? No. We'd say that we had discovered a water imposter. Since, therefore, water *must* be H_2O, WH is necessary, while also being a posteriori.

4.1 Evaluating KA

There is much that is right in this argument. Consider the substance that we bathe in, drink, etc. *If* that substance is in fact H_2O, it couldn't *fail* to be H_2O. And, just as Kripke says, you can't have a priori knowledge of the chemical composition of the liquid you bathe in.

But, contrary to what Kripke thinks, it doesn't follow that there are any necessary a posteriori truths. The problem is that, as before, Kripke confuses semantics with pre-semantic information.

Let us suppose that, when you were a child, your father pointed to the liquid in your drinking cup and said "that's water," and he said the same thing when pointing to the liquid in the bath you were about to step into, and so on. Kripke *correctly* says that:

(W_0) "water" refers to anything that is a clear liquid in drinking cups, bathtubs, etc.

is *false.* There could, in principle, be liquids in bathtubs, drinking cups, etc. that weren't H_2O but in day-today contexts were macroscopically indistinguishable from water, and those liquids wouldn't be water. But your father isn't *mis*-defining "water." For the correct interpretation of his definition is this:

(W_1) there is something x such that x is a clear liquid in drinking cups, bathtubs, etc., and "water" refers to x.

W_0 has the false consequence that "water" would refer to *anything* that is a clear liquid in bathtubs (etc). W_1 doesn't have this consequence. W_1 says that there is a liquid x that *as it happens* is in bathtubs and is clear, etc., but may or may not necessarily be so, and "water" refers to that liquid.

Let's look at the fact that you learn what "water" refers to throughW_1. W_1 *describes* the referent of "water." It says of this reference that, whatever it should turn out to be, it happens right now to be the clear liquid in bathtubs, etc. So it is *through* that description, or some other similar one, that you think about water. This doesn't mean that you incorrectly believe that *anything* that fits that description is water. It means that you think that what *in fact* fits that description—but might fail to—is what is referred to as water. So you know that there is something x that *in fact* fits that description (i.e., is a clear liquid in bathtubs, etc.) and that "water" refers to x.

Given it is by way of this sort of descriptive information that you know what "water" refers to, it's inevitable that:

(W_2) “water becomes ice when it’s very cold outside”

will convey the message to you that:

(W_3) there is something x that is a clear liquid in bathtubs, etc., and is referred to as “water”; moreover, x becomes ice when it’s very cold outside.

To be sure, W_3 is not W_2’s *literal meaning.* Kripke has made this clear. There is some clear liquid L such that, as it happens, L is in bathtubs, etc., and such that “water” refers to x and such that W_2’s literal meaning is:

(W_4) L turns to ice when it’s cold outside.

There’s nothing about bathtubs (etc.) inW_4.

For exactly analogous reasons, there is some liquid L that we bathe in, drink, etc., such that, for any property psi,

(W_5) ‹water has psi›

has for its literal meaning the proposition:

(W_6) L has psi,

even though it will convey the proposition that:

(W_7) there is something x that is a clear liquid in bathtubs, etc., and is referred to as “water”; moreover, x has psi.

In particular, it follows that:

(W_8) “water is H_2O”

will convey the message that:

(W_9) there is something x that is a clear liquid in bathtubs, etc., and is referred to as “water”; moreover, x is H_2O (i.e., consists of dihydrogen monoxide molecules),

even though there is something L such that W_8’s literal meaning is:

(W_{10}) L is H_2O.

W_9 is a posteriori. But it's also *contingent.* Water doesn't *have* to be in bathtubs. (There was water before there were bathtubs.) So we don't yet have anything that's both necessary and a posteriori. But, of course, it isW_{10}, not W_9, that is W_8's literal meaning. For the appropriate value of L, W_{10} is indeed necessary (just as Kripke says). But is W_{10} a posteriori?

We must make two preliminary points before we take a stand on this delicate question. First, we must distinguishW_{10} from the information *through which* it might be grasped. W_9 is an example of the sort of descriptive information through one graspsW_{10}. Since one grasps W_{10} through such propositions, it's easy to attribute their features to it. Since *they* are contingent, and we grasp W_{10} through them, it's easy to think that W_{10} is contingent.

The second point involves our outlining points that will be discussed more fully in Chapter 9. Consider the proposition:

(JS) John snores.

Under what circumstances is it true? The property of being identical with John must be instantiated. The property of being a smoker must be instantiated. And the property of being something that is and is also a smoker must be instantiated.

We can thus identify JS with a set S consisting of those three properties (viz. that of being identical with John, that of being a smoker, and that of being a thing that is John that is smoking). Since all the members of S are instantiated if and only if JS is true, we may identify JS's being true with its being the case that all the members of JS are instantiated. Thus, JS may be identified with a set of properties, and the fact that it's true may be identified with the fact that those properties are instantiated.

Let's apply this reasoning to L_{10}. That proposition is true if and only if anything that is an instance of the property of being identical with L is also an instance of the property of being identical with water. L *is* water. So the first property is the property of being identical with water. The property of being

identical with water *is* the property of being identical with H_2O. As Kripke himself stresses, it *isn't* the property of being the clear liquid in bathtubs, etc. No—it's the property of having a certain chemical composition, for something to have that composition is for it to consist of H_2O molecules. So if W_{10} is true the property of being with water is instantiated just in case the property of having a certain chemical composition, viz. that of being H_2O, is instantiated. Since the former is the property of having that composition, W_{10} is true just in case anything having that chemical composition has that composition. In other words, W_{10} is true just in case:

(W_{11}) anything that is H_2O is H_2O

which, though necessary, is a priori.

Once again, we find that Kripke's claim that he's found a necessary a posteriori truth involves his conflating *two* truths: one a posteriori and contingent, the other necessary but also analytic.

5.0 The causal vs. the descriptive theories of conception

According to Russell, "Socrates" is synonymous with a definite description; this might be "the great philosopher of antiquity who died of hemlock poisoning." To *understand* the word "Socrates," and thus to understand sentences of the form ‹...Socrates...›, one must obviously know the semantics of "Socrates." So one must know that "Socrates" is synonymous with that definite description. Consequently, anyone who uses the word "Socrates" with understanding is able to produce a description that applies to Socrates and Socrates alone. In general, if one knows the semantics of a referring term, then one knows some description that applies uniquely to the referent of that term. (A description "uniquely applies" to x if x uniquely satisfies it. So the description "president of the U.S. in 2009" applies to/is uniquely satisfied by Barack Obama.)

As we've seen, Kripke demolished this view. Consider *any* description of Socrates that Socrates uniquely satisfies (e.g., *greatest philosopher to have*

died of hemlock poisoning). One can understand ‹Socrates had psi› without knowing that a unique *greatest* philosopher to have died of hemlock poisoning had psi. The same thing *mutatis mutandis* holds of any description that singles out Socrates. It follows that “Socrates” isn’t synonymous with “the greatest philosopher to have died of hemlock poisoning” or with any other definite description and that the same thing *mutatis mutandis* is true of any other proper name.

5.1 The causal theory of reference

Rightly rejecting Russell’s view, according to which “Socrates” is a definite description that picks out Socrates, Kripke an alternative—to wit, “Socrates” refers to Socrates because a certain kind of *causal* relation holds between tokens of “Socrates” and Socrates himself (or, more exactly, states of affairs involving Socrates).

According to Kripke, somebody *saw* Socrates. (Let’s call that person “Smith,” even though that obviously wasn’t his name.) Thus, Socrates’ presence *caused* Smith to have certain visual experiences. Smith then gave a name to Socrates. (Since it doesn’t matter in this context that Socrates’ actual name was something of which “Socrates” is a distorted Anglicization, we’ll make the false but expedient assumption that Socrates was in fact named “Socrates.”) Smith passed that name along to people who did not themselves see, or otherwise sense-perceive, Socrates. In order for Smith to pass along that name to a given person, that person had to be causally connected to Smith, and *that* person had to be causally connected to anyone to whom he transmitted that word. Since, let us assume, I am the 521st installment in this sequence, I can refer to Socrates by saying “Socrates.” Thus, when I say “Socrates,” my words refer to Socrates because they are causally connected to him in a certain way.

In general, if an expression E refers to an object O, that is because tokens of E stand in a certain causal relation to states of affairs involving O.

5.2 The anti-descriptivist theory of conception

We’ve see that, for any property phi uniquely had by Socrates, one can refer to Socrates using the word “Socrates” *without* knowing that “Socrates” refers

to a unique instance of phi. Presumably, understanding ‹Socrates had psi›, for any property psi, involves having a concept of Socrates. Taken together, these points strongly suggest that one can understand "Socrates," and can therefore have a concept of Socrates, *without* knowing some truth of the form: *something x uniquely has phi, and "Socrates" refer to x.* This position is, of course, antithetical to the descriptivist conception of conception advocated by Russell.

5.3 The causal theory of conception

Because of Kripke's work, many philosophers accept both a causal theory of reference and an anti-descriptivist theory of conception. Many such philosophers believe that, since these two views of theirs are true, it is *de rigueur* to accept a *causal* theory of conception. According to such a theory, x has a concept of y if and only if y has certain sorts of effects on x.

David Kaplan (1968), himself an advocate of this view, made a point that helps bring out the raison d'être for it. Let X be a low-resolution, distorted photograph of Smith. Let Y be a high-resolution, undistorted photograph of somebody who isn't Smith but looks just like him. X can be thought of as a distorted, low-resolution, inaccurate, description of Smith, and Y can be thought of as a very good and accurate description of some Smith look-alike. The reason that, despite its shortcomings, X is a picture of Smith is that X is an *effect* of Smith (or, more accurately, of some situation involving Smith—light bounced off Smith and headed in the direction of a certain camera, etc.). And the reason that, despite its encoding a description that Smith satisfies to a tee, Y is not a photograph of Smith is that it lacks just this sort of causal connection to him.

Advocates of the causal theory of reference hold that concepts of Smith are to Smith what photographs of Smith are to him and that, for any x, concepts of x are to x what photographs of x are to x.

I reject this view. I accept Kripke's point that names aren't definite descriptions. And, with some qualifications to be stated forthwith, I accept Kripke's theory of reference. I agree that there is a causal *component* to conception, at least where spatiotemporal entities are concerned. (One's concept of a non-spatiotemporal entity does not, I believe, have such a

component.) But I agree with Russell that one's concept of Socrates *is* descriptive. There is, we will see, an ineliminable causal component to it. But that component is internal to a description that singles out Socrates. And I agree with Russell that one's concept of Socrates is descriptive and is expressed by a definite description. But, contrary to what Russell says, those definite descriptions aren't at all like the ones that Russell had in mind.

How can all of this be true? Simple: pre-semantics isn't semantics. For elucidation, read on.

5.4 The causal-descriptivist theory of conception

Imagine the following. There are 30 numerically different, but qualitatively identical men in a certain place. Your friend Bob goes up to one of them, put his hand on his shoulder, and says:

(1) This guy is named "Fred."

You thus know that:

(2) there is some individual x such that x is uniquely a person whom Bob just picked out; and "Fred" names x.

You thus know to whom "Fred" refers *only* because you know some true proposition of the form:

(3) something x uniquely has phi; moreover, "Fred" refers to x.

Your knowing who "Fred" refers to *also* depends on your having a causal connection to Fred. Your seeing Fred is important to your having this knowledge, and your seeing him involves your being at the receiving end of a causal process beginning with him. (x cannot see or otherwise sense-perceive y without being affected either by light signals bouncing off y or by other information-bearing processes originating with y.) But it's involved only because it's involved in your knowing some truth like (2)—some truth that *describes* the referent of the expression in question.

5.4.1 The causal-descriptivist theory of conception (continued)

Imagine that you are talking with your friends Smith and Jones. They suddenly starting talking about somebody named "Argo." They are

acquainted with Argo, but you are not. At no point in the conversation do they ever say anything like "Argo was my divorce lawyer" or "Argo was the best man at my wedding." They never explicitly or implicitly tell you anything of the form:

(4) something x uniquely has phi; moreover, "Argo" refers to x.

Suppose that, during this conversation, at time t, Smith says "Argo was a cad." Obviously the description ‹x is a cad› is indefinite, and many satisfy it. But on that occasion, there was, as you know, exactly one person who Smith was describing as being a cad, and you thus pick up knowledge of some claim along the lines of:

(5) on occasion t, there was exactly one person who Smith described as being a cad, and 'Argo' refers to x.

And, of course, (5) has (4) for its form.

So *merely* by virtue of being in a position to "pick up" use of the term "Argo," you are *ipso facto* in a position to know the truth of (1). Thus *merely* by virtue of being in a position to "pick up" use of the term "Argo" some existence-claim of the form: there is some x such that x uniquely has phi and "Argo" refers to x.

It is plain that what enables you to pick up the term "Argo" is your knowledge of some claim of the sort just described. Your picking up a term like "Argo" involves your hearing or reading a sentence of the form ‹...Argo...›. It thus involves your encountering a case of somebody's ascribing a property to Argo. (This is true even if you read such sentence: you are *still,* in effect, encountering somebody's using that sentence to make such a statement.)Whenever you hear a token of ‹Argo has phi› you are thereby made aware of the fact that there is somebody x such that, *on that occasion,* the property of being a phi is being ascribed to x and x alone; and you are also being made aware of the fact that, on that occasion, "Argo" is being used to refer to x. On any such occasion, therefore, you are being apprised of an existence-claim that Argo uniquely satisfies, even though no such claim is being overtly stated.

We must be careful to distinguish the following claims:

(a) Exactly one person has phi.

(b) Exactly one person is being *described,* on some particular occasion, as having phi.

There are many cads. But there are many occasions where only one person is being *described* as a cad. Your learning who "Argo" refers to doesn't involve your learning that:

(6) exactly one person x is a cad, and "Argo" refers to x.

You couldn't possibly "learn" (6), since it's false, given that there are many cads. Your learning who "Argo" refers to involves your learning that, on some occasion, somebody x is being **described** by some specific person (Smith) as a "cad."

What these remarks show is that, even though names are merely labels, we can *attach* them to their referents only because we know of descriptions that single those referents out. Semantically, "Argo" and "Socrates" are mere labels. But *pre*-semantically, they are descriptions. The information that one must use to affix them to their referents is descriptive in nature, and so is the information one must work through to interpret utterances containing occurrences of them.

5.5 The problem of forgotten start-up descriptions

One typically forgets the existence-claim through which one added a name to one's lexicon. One typically forgets the *first* time one heard "Shakespeare" or "Aristotle"; one often forgets one's *first* meeting of a person. Given this, one is tempted to say:

> You don't necessarily associate *any* uniquely individuating description with "Aristotle" or "Smith" or "Jones." Such a description was needed to *enable* you to refer to those people, and to think about them. But once that ability is in place, the description drops out. At that point, there can be *pure* thought about the entities in question. You can *just* think about Aristotle and Smith; you don't have to think about them *through* a uniquely individuating description.

Remember what we said earlier about "Fred." Your adding "Fred" to your lexicon involves your learning some truth of the form: *somebody x*

uniquely has phi and "Fred" names x.

Bearing this in mind, suppose that you and a friend are at a party. You see somebody wearing a bowler hat, talking animatedly. At time t, your friend points to that person and says "*that* is Fred." Your adding "Fred" to your lexicon involves your knowing some existence-claim along the lines of:

(E_1) there is now somebody x who I am sense-perceiving and who is wearing a bowler hat; and "Fred" names x.

Let us suppose that, a minute or so after "Fred" is ostensively defined for you, you go up to him and start talking. Now you have new information about Fred; you have a new uniquely individuating description of him. There are a number of reasons for this, the mere passage of time being one of them. At time t, E_1 was your uniquely individuating description. But now that time has passed, E_1 has been replaced with:

(E_2) there is somebody x whom, a minute or two ago, I was sense-perceiving and attending to, and who was then wearing a bowler hat, and "Fred" names x.

Notice that E_2 "interlocks" with E_1. The uniquely individuating description changes. But the new one is related to the old one in such a way that some one object uniquely satisfies both of them. The new one thus "inherits" the verifier of the old one. In a moment we will say more about this process of inheritance or transmission.

Things other than the mere passage of time, will lead to E1's being replaced. Let's suppose that you are now engaged in a conversation with Fred. You can see him more clearly than before. (Although, when looked at from afar, he looked healthy and strong, you now see that he is puffy and out of shape.) He is talking about the national debt. As a result, you generate a *new* uniquely individuating description of Fred, this one being given by an existence-claim along the lines of:

(E_3) there is somebody x such that, a minute or two ago, I was sense-perceiving x while standing several feet away from him, and such that, at the present time, I am listening to his puffy and out of shape self talk about the

national debt; moreover, “Fred” names x.

Notice that E_3 and E_2 interlock; E_2 transmits its verifier to E_3.

Now that you have E_3, you can *drop* E_2 and E_1. You can completely forget them; and that will not jeopardize your ability to think about Fred. Supposing that a moment has passed and that you have stopped talking with Fred, the mere passage of time replaces E_3 with a slightly different claim:

(E_4) there is some x such that, a moment ago, x was talking about the national debt, and x looked puffy and unhealthy, and “Fred” names x.

Given that you now know E_4, you can forget each of E_1–E_3 without ceasing to be able to think about Fred. And, so long as you are able to think about Fred, this will continue to be the case.

One’s ability to think about, and refer to, something can only be *started up* by one’s acquiring knowledge of the right kind of existence-claim. Typically, that specific claim is forgotten. But it is replaced by another interlocking existence-claim. And that second one will, in its turn, be replaced by a third; and so on. Thus, an integral part of one’s ability to think about or refer to an external object is one’s having knowledge of an existence-claim that it uniquely satisfies; i.e., it is my having a uniquely individuating description of him.

None of this belittles the role that causal connections play in conception and reference. I am able to think about, and refer to, Fred precisely because I saw him: and seeing something involves being causally connected to it. But *how* is the causal connection involved? What enables me to think about Fred is my knowledge of some existence-claim that describes him: *there is some guy whom I am now seeing and attending to, who is wearing a bowler hat, and “Fred” names that guy.* The causal connection is involved in my seeing Fred; and my seeing Fred, in its turn, is involved in my having a uniquely individuating description of him. So what is doing the work is my knowledge of a uniquely individuating description of Fred; the causal connection, though essential, has a role that is subordinate and internal to my knowledge of that description.

Here we need to very clear about a certain distinction: the distinction between *literal meaning* and *cognitive content.* "Fred" is just a label. The following statement is true:

(F_1) There is somebody x who I met on such and such occasion and who looked pale and talked about communism; and the semantic rule for "Fred" is: "Fred" names x.

The following statement is false:

(F_2) The semantic rule for "Fred" is this: There is somebody x who I met on such and such occasion and who looked pale and talked about communism; and "Fred" names x.

The following embodies even more falsehoods than F_2:

(F_3) The semantic rule for "Fred" is this: for any predicate ‹...x...›, ‹...Fred...› means: There is somebody O who I met on such and such occasion, and who looked pale and talked about communism and...O...

There is some x such that the literal meaning of

(*) "Fred looked puffy and pale"

is simply:

(*LM) x looked puffy and pale.

But remember that one knows what "Fred" refers to only insofar as one knows some claim of the form: *somebody x uniquely has phi and "Fred" refers to x.*

Tyler Burge (1941–) overlooked these points. The result of his doing so was a doctrine known as ***content externalism*** that, in order to be kept from crashing, has had more epicycles added to it than just about any theory in human history. It is to this dark chapter of intellectual history that we turn.

Chapter 9

Putnam's Insight and Burge's Blunder:

Semantic Externalism vs. Content-Externalism

1.0 Is linguistic meaning a strictly psychological notion?

In *Naming and Necessity,* Kripke decisively showed that proper names are not definite descriptions. In doing this, Kripke was making a specific point concerning the semantics of a specific class of expressions. It might therefore seem that what Kripke was doing in NN, though non-trivial, was too narrow to be of great philosophical importance.

But implicit in what Kripke said about proper names are many deep and important principles concerning the very nature of meaning and the relationship between language and thought. In 1975, Hilary Putnam published an article titled *The meaning of "meaning*" (henceforth, "MM") in which he explicitly stated one of these principles and cogently argued for it. What he showed is that what we think does not, by itself, establish what our words and sentences mean.

It's obvious that what we think *largely* determines what it is that our expressions mean. The word "snow" could refer to anything; it would be easy to create a language in which it referred, not to so snow, but to Julius Caesar. (Let's create such a language right now. Let L be a language that is just like English except that, in L, "snow" refers to Julius Caesar and "Julius Caesar" refers to snow. L is a perfectly good language—it's as expressively powerful a language as English.) It's obviously at least partly because of what some people were thinking at certain times that "snow" refers to snow, and not to Julius Caesar.

But Putnam made it clear that there is an important *non*-psychological dimension to linguistic meaning. Given only that, at any given moment, person X is psychologically and behaviorally indistinguishable from person

Y, it does *not* follow that the words leaving X's mouth have the same meanings as the words leaving Y's mouth. So, supposing that at any given moment X is having a given thought if and only if Y is having an exactly similar thought, and that X is producing a given utterance if and only if Y is producing an exactly similar utterance, it does *not* follow that what is meant by X's utterance coincides with what is meant by Y's. Even though the psychological state that gave rise to the one utterance is just like the psychological state that gave rise to the other, the meaning of the one may diverge from that of the other.

1.1 Putnam's argument

Let Twin-Earth be a planet in some distant galaxy that satisfies the following conditions. Given any person x on Earth, there is a person y on Twin-Earth who is an atom-for-atom duplicate of x. At any given moment, y has a given thought if and only if x has that thought; y has a given feeling if and only if x has that very feeling. y acts a certain way if and only if x acts in an exactly similar way; y makes a certain noise if and only if x makes an exactly similar noise. Thus, Earth people are psychologically and behaviorally exactly like Twin-Earth people. They think, feel, do, and say the very same things at the very same times.

It is merely coincidence that Earth people are in lock-step with Twin-Earth people. There is no causal interaction between Earth people and Twin-Earth people. No messages are sent from one planet to the other; nobody on either planet in any way influences anyone on the other.

Even though Earth and Twin-Earth are extremely similar, there is an important difference between them. On Twin-Earth, the liquid in bathtubs, drinking glasses, lakes, etc., isn't H_2O—it's XYZ. ("XYZ" is an abbreviation for some long chemical formula.) With that qualification, Twin-Earth is, at any given moment, exactly like Earth.

Let's continue our story. Smith is a psychologically normal, English speaking, adult Earthling who lives in the 1600s. One day, after doing back-breaking work for several hours, Smith desperately wants to drink some cool water. Addressing his companion, Smith says:

(SE[163]) "When people are thirsty, they find it more refreshing to drink

cool, glacial water than warm, swamp water; and if there were no clean water, people couldn't survive."

This utterance takes place at time t.

Important point: SE doesn't refer to a sentence. It refers to a specific *utterance* of a sentence. It refers to the specific burst of noise produced by Smith at time t. (Different people can *utter* a given sentence; different people can *utter* the sentence uttered by Smith (viz. "when people are thirsty, they find it more refreshing to drink cool, glacial water than . . .")

Bear in mind here that, in Smith's time, nothing was known about the chemical composition of water. People obviously had thoughts about water; they wanted to drink it; bathe in it, and so on. But they didn't know that it was H_2O; they didn't even grasp the concepts needed to grasp the concept of an H_2O molecule.

On Twin-Earth, there is somebody who, at any given moment, is psychologically and behaviorally just like Smith. (This follows from what we said earlier.) Let Twin-Smith be that person. At any given time, Twin-Smith makes the very noises as Smith and, in so doing, expresses the very sentiments. (This too follows from our earlier points.) Thus, at time t, Twin-Smith says:

(TS[164]) "When people are thirsty, they find it more refreshing to drink cool, glacial water than warm, swamp water; and if there were no clean water, people couldn't survive."

Important point: TS doesn't refer to a sentence. It refers to a specific *utterance* of a sentence—to the specific burst of noise produced by Twin-Smith at time t.

Question: Does SE have the same meaning as TS? I am not asking whether Smith and Twin-Smith coincide in respect of what *they* mean by their respective utterances, but whether SE and TS coincide in respect of what *they* mean. The question doesn't concern what Smith and Twin-Smith are trying to convey through their respective utterances; it concerns what those utterances themselves mean.[165]

SE and TS *don't* have the same meaning. They have different

propositions for their respective meanings, and the one proposition is true in circumstances where the other is false, and false in circumstances where the other is true.

"Water" refers to the substance whose chemical composition is H_2O. That's a simple fact. A thoroughly competent speaker of English can easily fail to *know* that "water" refers to that substance. Newton didn't know it, even though he was an exceptionally intelligent person who spoke perfect English. But that doesn't matter. Here on Earth, there is a certain liquid x such that x is found in bathtubs, drinking glasses, lakes, and oceans, and such that that is referred to as "water." That substance was referred to as "water" for many hundreds of years before anything was known about chemistry or, in particular, about the chemical composition of that substance. When, finally, that substance was analyzed, it turned out to consist of H_2O molecules. Thus,

(WAT_1) "there is some substance x such that x is found in bathtubs, drinking glasses, lakes, and oceans, and such that x consists of H_2O molecules, and such that: '*water' refers to x.*"

Of course, we now know that the substance referred to as "water" is H_2O. And if we were to discover some substance that consisted of XYZ, not H_2O, but was otherwise just like water, at least as far as the conditions that are likely to obtain on the surface of the Earth are concerned, that other substance wouldn't be H_2O. The chemists who discovered XYZ would *not* come to the conclusion that water is *sometimes* H_2O and is sometimes XYZ. No—they'd say that water is *always* H_2O, but that a different substance, namely XYZ, is extremely water-like, at least within certain horizons.

When they weren't referring to XYZ by its chemical name, chemists would *not* refer to it as "water." Maybe they'd refer to it as "pseudo-water" or as "fool's water"—but not "water." This practice of theirs wouldn't be a case of hair-splitting pedantry. If chemists *did* refer to this other substance as "water," their doing so would violate beliefs that are deeply embedded in common sense. Even the most philistine, intellectually narrow human being desperately wants real diamonds, not zirconium imitation diamonds, and thus

knows that being diamond-*like* isn't the same thing as being a diamond. Similarly, everybody knows that being water-*like* isn't the same thing as being real water. And even if some people don't know it, they'd quickly learn it, no matter how scientifically uninformed they were, if the distinction between water and XYZ became a financially significant one, like the distinction between diamonds and chunks of zirconium paste that resemble diamonds.

For the same reasons *mutatis mutandis,* on Twin-Earth "water" would refer to XYZ, not H_2O. And it would *always* have referred to XYZ—even before scientists on their planet discovered its chemical composition.

Before proceeding, let's deal with an objection that is likely to be heard:

> An English-speaking physical chemist obviously knows a lot about water that other English speakers don't. But what the physical chemist knows that other English speakers don't has nothing to do with the English language or, therefore, with *semantics.* Given only that somebody doesn't have a PhD in physical chemistry, it doesn't follow that they have any *linguistic* deficits—that they don't *really* know the identities of the semantic rules governing expressions like "water," "gold," "lead," etc. This is all quite uncontroversial.
>
> But it's inconsistent with the view you (speaking on Putnam's behalf) are advocating. You say that, for something to be water, it isn't enough that it have certain macro-characteristics: it must have a specific chemical composition—it must be H_2O. Given that, as a matter of semantics, "water" refers to water, what you're saying entails that, as a matter of semantics, "water" refers to H_2O. But if that were right, one couldn't master the English language without knowing the chemical composition of water or, by obvious extensions of this argument, the chemical composition of gold or aluminum. But that's preposterous.

One doesn't have to know anything about chemistry to have a complete mastery of English (or any other natural language), and it would be preposterous to deny that. But the view I'm describing doesn't have that consequence. There is some individual x such that the semantic rule for the

expression “Barack Obama” is to the effect that:

(BO) “Barack Obama refers to x.”

BO pairs off a name with an individual. If one doesn’t know which individual that is, then one doesn’t know the semantic rule for “Barack Obama” and, therefore, one doesn’t know all of the semantic rules constitutive of the English language in 2009.

But there are infinitely many ways of singling out any given entity. This is a consequence of the fact that, as a matter of logic, any given object uniquely satisfies infinitely many different descriptions. Let P be the region of space you were occupying the instant you began reading this sentence, and let t be that instant. You uniquely satisfy each of the following predicates: “thing that occupied p one minute before time t + one minute,” “thing that occupied p two minutes before time t + two minutes,” and so on. We’ll henceforth abbreviate the first of these predicates as “P_1,” the second as “P_2,” etc.

For argument’s sake, let’s suppose that your name is “Jack Jones.” There is some x such that x uniquely satisfies each of P_1, P_2, P_3, etc., and such that, in order to the know semantic rule governing that expression, one must know that:

(JJ) “Jack Jones refers to x.”

If Smith knows that:

(JJ_{75}) there is some x such that x uniquely satisfies P_{75} and such that: “*Jack Jones” refers to x*

and Jones knows that:

(JJ_{143}) there is some x such that x uniquely satisfies P_{143} and such that: “*Jack Jones” refers to x*

Smith and Jones are, to that extent, in complete semantic accord. There is some one x such that x satisfies each of P_{75} and P_{143}, and such that, given any property psi, when Smith says:

(JJ_1) ‹Jack Jones has psi›

he is affirming that

(JJ_2) x has psi,

and such that when Jones utters JJ_1, he too is affirming that x has psi. Further, it is each individual's *intention* in uttering JJ_1 to affirm JJ_2. What each wishes to do is to attribute psi to some person—nothing more. Neither wants to say anything about his way of thinking about that person. Neither wishes to say of P_{75} or P_{143} or any other property that it is uniquely instantiated or that all its instances are psis.

Points exactly similar to those just made about "Jack Jones" hold with respect to "water." At time t, bathtub B is filled with pure water, on which your rubber ducky is floating. Your mother points to that B and says:

(*) The liquid in the tub is known as "water."

A few days later, at time t*, you burn your finger in scalding water. Given that water is H_2O, it follows that each of the following statements is true:

(1) there is some liquid x such that, at time t, x is what your rubber ducky is floating on; moreover, *"water" refers to x.*

(2) there is some liquid x such that, at time t, x is what your bathtub is full of; moreover, *"water" refers to x.* (3) there is some liquid x such that x and x alone is H_2O, and such that *"water" refers to x.*

(4) there is some liquid x such that, at time t*, x and x alone is what scalded your finger; moreover, *"water" refers to x.*

(5) there is some liquid x such that x and x alone is H_2O and such that, at time t*, x and x alone is what scalded your finger; moreover, *"water" refers to x.*

And so on.

To know what "water" refers to, it is necessary only that you know *some* truth of the form:

(n) there is some substance x such that x uniquely has property phi; *"water" refers to x.*

Since water uniquely has each of infinitely many different properties, n is true for each of infinitely many values of phi. Where any given one of these truths is concerned, knowing it is sufficient, but not necessary, for knowing what "water" refers to and, therefore, for knowing the semantics of that expression.

Since n is a sentence-schema, not a sentence proper, it does not itself express any one of these truths. But in what follows, take n to stand for an arbitrary one of them.

Only the italicized part of n describes a semantic rule. The rest describes the information through which one grasps that rule. Anyone who knows the semantics of water knows this (though they might not recognize what it is that they know in the statements just made).

Let's tell another story about Smith and Jones. Both are competent speakers of English. Smith knows that:

(7) there is some liquid x such that x is what is in bathtubs and drinking glasses; moreover, *"water" refers to x.*

Jones, who is more scientifically minded than Smith, knows that:

(8) there is some liquid x such that x is H_2O and such that creatures with gills can breathe when immersed in x and creatures with lungs cannot; moreover, *"water" refers to x.*

There is some liquid x such that individual's sole intention in saying

(w_2) "water covers most of the Earth's surface"

is to say:

(w_3) x covers most of the Earth's surface.

Neither wishes to affirm either that:

(w_4) there is some liquid x such that x is what is in bathtubs and drinking glasses, and such that "water" refers to x; moreover, x covers most of the Earth's surface,

or that

(w_5) there is some liquid x such that x is H_2O and such that creatures with gills can breathe when immersed in x and creatures with lungs cannot; moreover, x covers most of the Earth's surface.

Each individual's intention in uttering w_2 is to ascribe a certain property to a certain substance. They have the same substance in mind. They wish to attribute the same property to it. Neither wants to say anything about how he thinks about that substance. Neither wants to attribute any property to anything other than that substance. Thus, what Smith means when he says "water covers most of the Earth's surface" is exactly what Jones means when *he* utters the same sentence. And so long as each knows that the other agrees with him about what it is that "water" that refers to, any difference between them in respect of how they think about its referent will in no way jeopardize the one person's ability to understand the other. To put it more formally: provided that, for some x, each believes that "water" refers to x and that each knows that the other believes this, the differences between these two individuals in respect of how they think about that substance won't

jeopardize their ability to communicate with each other.

1.2 Putnam's semantic externalism (continued)

These points make it clear why it is that Smith and Twin-Smith, though driven to make the very same noises by the very same internal conditions, are nonetheless affirming different propositions. Let's suppose that when Smith was a child, his mother pointed to the liquid in the tub in which Smith regularly bathed, and which Smith recognized as the liquid that he had to drink in order to live, and said: "That is water." The message that is communicated to Smith is along the lines of:

(8) There is some liquid x such that x is what I (Smith) bathe in and must also drink in order to survive and "water" refers to x.

Given that the liquid indicated by Smith's mother is, of its very nature, H_2O, it follows that when, at time t, Smith utters SE, his utterance is true exactly if:

(SE*) "When people are thirsty, they find it more refreshing to drink cool, glacial H_2O than warm, swamp H_2O; and if there were no clean H_2O, people couldn't survive.

Because he lives hundreds of years before the birth of modern chemistry known, Smith has no idea what it would even mean to identify water with H_2O. Moreover, *nobody* who is alive at the time has any such idea. Nonetheless, the statement he is making does concern H_2O. This shows that:

(PTM) the meanings borne by our utterances are not determined solely by the thoughts we have; there is a strictly non-psychological, strictly environmental dimension to linguistic meaning.

Given what we know about Twin-Smith and Twin-Earth, it follows that, when Twin-Smith utters TS, what he is saying is true exactly if:

(TS*) "When people are thirsty, they find it more refreshing to drink cool, glacial XYZ than warm, swamp water; and if there were no clean XYZ, people couldn't survive."

And this is obviously consistent with Putnam's thesis (henceforth "PTM").

The doctrine that PTM is true is known as *semantic externalism.* It is so-called because, according to PTM, the determinants of the semantics of our expressions are (in some cases, in some respects) external to us.

2.0 Burge's blunder—the (mis)step from semantic to content externalism

PTM is a deep, important, and original insight, and Putnam deserves sole credit for it. And unlike many of the philosophers who were influenced by his important article, Putnam's views as to what PTM does and does not imply were correct. In MM, Putnam says that Smith's *words* don't have quite the same meanings as Twin-Smith's *words,* even though Smith and Twin-Smith have the same *thoughts.*

Under pressure from other philosophers, especially Tyler Burge[166], Putnam[167] later rejected this correct and profound insight, and he came to accept Burge's position, which is that, not only do Smith's *words* have different *meanings* from Twin-Smith's, Smith's *thoughts* don't have the same *contents* as Twin-Smith's.

Burge's work has been enormously influential. It didn't just influence Putnam, but analytic philosophy as a whole. Thanks largely to Burge's work, it came to be widely held for 20 years or so that Smith and Twin-Smith have different *thoughts.* This must be understood aright. Nobody denies that Smith and Twin-Smith have numerically different thoughts. Burge is making the much less innocuous claim that what Smith is thinking is different from what Twin-Smith is thinking. In Burge's view, just my thought that it's hot during the summer has a different content from your thought that it's cold during the winter, so, according to Burge, the thought that Smith expresses in saying, for property psi, "water has psi" is different from the thought that Twin-Smith expresses by means of that same utterance.

This interesting chapter in the history of analytic philosophy began in 1979, four years after Putnam published MM, with the publication of Burge's article "Individualism and the mental" (henceforth, "BI"[168]). Here is what Burge argues. There is a planet, which we'll call "Twin-Earth," that is exactly like Earth, with one qualification to be stated in a moment. The events on

Twin-Earth are in lockstep with those on Earth. So given that earthling Smith sneezes at a certain time, Twin-earthling Twin-Smith sneezes in the exact same way at the exact same time.

Given what we've said, it follows that English is spoken by many people on Twin-Earth. Twin-Smith is one of these people. On our planet—on Earth, in other words—"arthritis" refers to a disease exclusively of the joints. On Twin-Earth, "arthritis" refers to a malady not *only* of the joints, but also of certain bones (as opposed to the cartilaginous joints linking bones to one another). So, on Twin-Earth, "arthritis" doesn't refer to the same disease that it refers to on Earth. It refers to a similar, but distinct ailment. Let us refer to that ailment as "twin-arthritis."

On Earth, Smith regards the sentence:

(AU[169]) "arthritis is unpleasant"

as expressing a truth. And because he personally suffers from arthritis, he often utters AU. (In this context, it is to be assumed that Smith is speaking sincerely, is cognitively competent, etc., and these same assumptions are to be made about Twin-Smith).

Like Smith, Twin-Smith regards AU as expressing a truth. But, according to Burge, the thought underlying Twin-Smith's utterance of AU is different from the thought underlying Smith's utterance of it. Twin-Smith is thinking about twin-arthritis, not arthritis, and Smith is thinking about arthritis, not twin-arthritis. In a moment, we'll evaluate this important claim.

Burge concludes from this that two people who are atom-for-atom duplicates of each other, each a perfect physiological and psychological clone of the other, may have thoughts with different contents. Twin-Smith is thinking *Descartes* (or Twin-Descartes, rather) *was a great philosopher,* whereas Smith is thinking *Ted Danson is a great philosopher,* notwithstanding that Smith and Twin-Smith are, by Burge's hypothesis, psychologically exactly the same.

Is Burge right? First of all, he's right that, on Twin-Earth, "arthritis" doesn't *refer* to arthritis, and instead refers to twin-arthritis. This is precisely what Putnam showed. But is Burge right that Smith and Twin-Smith are *cognitively* different?

No. How does Smith learn what "arthritis" refers to? Here's a possible answer (and it's representative, in the relevant respects, of all the other possible answers). Addressing Smith, so and so says:

(A_1) "the unpleasant inflammation in your knees is due to arthritis; arthritis can do terrible damage to one's joints."

A_1 is to the effect that:

(A_2) there is some x such that the swelling in your knees is due to x and such that x can do terrible damage to one's joints; moreover, "arthritis" refers to x.

For now familiar reasons, it follows that when Smith hears someone say:

(A_3) "arthritis has psi,"

the message that is communicated (not literally meant) is:

(A_4) There is some x such that the swelling in your knees is due to x, such that x can do terrible damage to one's joints, and such that "arthritis" refers to x; moreover, "arthritis" refers to x.

Remember that Twin-Smith's life precisely parallels Smith's. It therefore follows that, when Twin-Smith hears someone utter (A_3), is the message that is communicated (not literally meant) is (A_4).

Thus, contrary to what Burge says, Smith's thoughts *do* have the same content as Twin-Smith's. It's true that, for the reasons given by Putnam, the *sentences* they use to express those thoughts don't always have the same literal meanings. But unless one countenances the obviously false assumption that what we think coincides exactly with what our expressions mean, it doesn't follow that Smith's thoughts have the same contents as Twin-Smith's .

2.1 Literal meaning vs. cognitive content

An utterance on Earth of "arthritis is painful" *does* have a different literal meaning from such an utterance on Twin-Earth. There is some malady x, namely arthritis, such that x only afflicts the joints and such that an utterance *on Earth of:*

(A_5) "arthritis is painful"

is true just in case:

(A_6) x is painful,

it being irrelevant whether some disease that afflicts things *other* than the joints is also painful. And there is some y, namely twin-arthritis, such that y does not afflict only the joints and such that an utterance on *Twin*-Earth of

(A_5) "arthritis is painful"

is true just in case:

(A_6) y is painful,

it being irrelevant whether some disease that afflicts *only* the joints is also painful. But Burge is wrong to conclude from these facts that Smith and Twin-Smith have different thoughts.

2.2 A second Burge-style argument[170]

Let W_1 be the actual world and let W_2 be some other possible world. In each of W_1 and W_2, my friend Max has an identical twin Gary. Gary and Max are cosmetically, physiologically, and behaviorally completely and totally indistinguishable. To know that one is seeing Max as opposed to Gary, or *vice versa,* one has to rely on ulterior knowledge (e.g., that Gary is now in Finland and that he couldn't possibly be the person with whom one is now lunching in New York).

In W_1, at time t, my friend Sally and I are 10 feet away from Max, and I'm looking right at him. In W_2, at time t, Sally and I are 10 feet away from Max's identical twin, Gary. Otherwise, W_1 at t and W_2 at t are indistinguishable (at time t). Given only the information available to me by virtue of having the perception involved in my seeing Max, it is not even a theoretical possibility that I might encounter some evidence or have some experience on the basis of which I could rationally conclude that I was in the one world as opposed to the other.

Moreover, the experiences that I have had in W_1 up until time t have been qualitatively identical with those that I've had in W_2 up until time t. So neither my current experiences nor my memories nor the two combined could possibly enable me to determine whether I'm in the one world as opposed to the other.

It is to be assumed in what follows that all my utterances are sincere—they express beliefs that I actually have.

In W_1, while pointing at Max, I tell Sally:

(U_1[171]) "That entity is wearing a plaid shirt."

(In due course, it'll be clear why I'm using the word "entity" as opposed to

individual.) In W_2, while pointing at Gary, I tell Sally:

(U_2) "That entity is wearing a plaid shirt."

U_1 and U_2 are specific *utterances.* So U_1 is an utterance that occurs in Max's presence in W_1, and U_2 is an utterance that occurs in Gary's presence in W_2. In each world, I know that I'm seeing *either* Gary *or* Max, but in neither world do I know which of the two I'm seeing.

In W_1, there is some individual x, namely x, such that such x ≠ Gary and such that U_1 is true exactly if:

(MP[172]) x is wearing a plaid shirt.

In W_2, there is some other individual y, namely Gary, such that y ≠ Max and such that U_2 is true exactly if:

(GP) y is wearing a plaid shirt.

MP is true just in case *Max* is wearing a plaid shirt, it being irrelevant whether Gary is wearing one. And GP is true just in case *Gary* is wearing a plaid shirt, it being irrelevant whether Max is wearing one. Neither entails the other; neither entails the negation of the other. MP and GP are different propositions.

U_1 expresses what I'm thinking in W_1, and U_2 expresses what I'm thinking U_2. U_2s ' truth-conditions coincide with those of the thought I'm having in W_1, and are therefore given by MP; and U_2's truth-conditions coincide with those of the thought I'm having in W_2 and are therefore given by GP. Thus, the belief expressed by U_1 has a different content from the belief expressed by U_2.

Of course, in W_1, my visual experience results from light bouncing off Max, not Gary; and in W_2, it results from light bouncing off of Gary, not Max. But apart from that fact, there is absolutely nothing that at that time distinguishes my situation in the one world at time t from my situation in the other world. My intracranial situation in W_1 is just like my intracranial situation in W_2.

But we just saw that the content of what I'm thinking in W_1 at t is different from the content of what I'm thinking in W_2. This difference must derive from the fact that, in W_1, my mental state has a certain external cause (light bouncing off of Max), whereas in W_2 it has a different external cause (light bouncing off of Gary).

2.3 Burge's failure to distinguish what we know directly from perception from what we infer from perception

The conclusion of this argument is false, and there is thus something wrong with the argument itself. The problem with it is that it assumes that the relationship between linguistic meaning and cognitive content is much tighter than it actually is.

Let's suppose the following. I'm in some world W_3 (not identical with W_1 or with W_2). I'm with my friend Sally. I bet her $10 that within the next 24 hours we'd see somebody wearing a plaid shirt. She doesn't know, or know of, either Max or Gary. W_3 is exactly like each of W_1 and W_2 *except* that, in W_3, the thing standing in front of me isn't Max or Gary—it's a robot that looks exactly like both of them. I don't (yet) know that such a robots exists. I'm friends with Max and Gary. I happen to know that Gary is out of the country. I also know that Max is somewhere in the vicinity. So, given that I'm looking at something that looks exactly like Max, I believe that I *am* seeing Max. I'm with my friend Sally. I see that the thing I'm beholding is wearing a plaid shirt, and I say:

(U_3) "That entity is wearing a plaid shirt."

Let VR[173] be the visual perception of which my seeing this thing consists. VR is utterly high resolution and veridical. I'm seeing the thing before me, whatever it is, *as* it is. Given that this thing looks exactly like Max and Gary, my seeing it as it is involves my having a perception exactly like the one that I'd have in that very situation if *ceteris paribus,* it was Max or Gary I was seeing.

Thus, VR does not itself tell me who it is that I'm seeing. Given only what VR tells me, that thing could be Max or Gary or some third person (or thing). VR does tell me that I'm seeing somebody with such and such characteristics, among them that of wearing a plaid shirt.

I'm wrong about who (and what) that entity is. But that doesn't affect the belief I'm trying to get across. I don't care who or what that entity is, and neither does Sally. All we care about is whether it's wearing a plaid shirt. I correctly believe that it is wearing a plaid shirt, and that's the belief I'm trying to get across by uttering U_3. Therefore, the belief embodied in U_3 is correct.

This has two very important consequences, each of which is incompatible with Burge's position. Consequence #1: There is no particular x such that VR tells me that it's x I am seeing. This follows from the fact, just noted, that VR does not itself apprise me of the identity of the thing that I'm seeing.

Consequence #2: There is no particular x such that the belief embodied in U_3 is correct only if x is plaid shirt. This follows from the fact, earlier noted, that the belief embodied in U_3 is correct as long as the thing we're seeing is wearing a plaid shirt, it being irrelevant who or what that thing is.

Suppose that Robo-Max is the entity that I'm seeing—that he's the object of VR. Given only what VR is telling me, I could be seeing any number of different individuals. What VR tells me is that there is *somebody* having such and such characteristics—that there is *somebody* standing in a certain place and wearing a plaid shirt. So the content of VR is existential. It tells me that:

(CVR[174]) There is some entity x such that x has such and such characteristics, such that x is standing in a certain place and such that x is wearing a plaid shirt.

Obviously, CVR is at least one of the things that, on the basis of VR, I believe to be true. And it's the relevant belief in this context, since it's the one embodied in U_3.

Supposing that VM is the perception I'm having in U_1 of Max, and that VG is the perception I'm having in U_2 of Gary, it follows that the belief embodied in U_3 has the very same content as the belief embodied in U_1 as well as the belief embodied in U_2. Remember that, from an experiential standpoint, no two of these situations are distinguishable, even in principle.

Let's state as clearly and explicitly as possible how Burge goes wrong. Burge says that what VR is telling me is *different* from what VM is telling me and also from what VG is telling me. Burge thinks that there is some x, namely a certain robot, such that VR is correct just in case:

(bx) x is wearing a plaid shirt.

And Burge thinks that there is some y, namely Gary, such that VG is correct just in case:

(by) y is wearing a plaid shirt.

Finally, Burge thinks that there is some z, namely Max, such that VM is correct just in case:

(bz) z is wearing a plaid shirt.

Of course, none of these three propositions is equivalent with other; none so much as entails either of the others. So if Burge is right, then the proposition encoded in my visual perception in W_3 is different from my visual perceptions in W_1 and W_2, and the propositions encoded in the latter two perceptions are different from each other.

But Burge is not right. Given only what VR is telling me—in other words, given only what I know on the basis of VR *without* the help of pre-existing, background knowledge derived from some source other than VR—I can't possibly know that I'm seeing robot x as opposed to person y as opposed to person z as opposed to Martian w. The content of my perception on W_3 is, quite obviously, neutral with respect to the identity of the object of that perception, and its content is therefore, in that respect, general and, more specifically, is given by CVR.

Russell (1927) makes this point very clearly:

> Prejudice tells you that you see the *same* table on two different occasions; you *think* that experience tells you this. If it really were experience, you could not be mistaken; yet a similar table may be substituted without altering the experience . . . there is nothing to show that one identical entity causes the two sensations.[175]

The message embodied in VR goes through regardless of who it is that I am seeing to be wearing a plaid shirt. That identity-independent message is embodied in the belief underlying my utterance of "that guy is wearing a plaid shirt," and it's embodied in that utterance. Therefore, Burge is simply wrong to identify the

content of that belief with bx. For exactly similar reasons, Burge is just plain wrong to identify bx (by)with what VM (VG)tells me and with the belief embodied in U_1 (U_2).

Of course, there are differences between those situations. In the first, it's Max who's standing in front of me; in the second, it's Gary; in the third it's a robot. But these differences don't lead to any experiential differences, since those three things all look exactly the same. In none of these worlds can the content of my visual perception, taken by itself, register that it's Max, Gary, or some third thing that's in front of me. Given that

CVR is the content of my visual perception in W_3, it follows that CVR is also the content of my perception in U_1 and also of my perception in U_2. (The content of my visual perception in the one world coincides with the content of my visual perception in each of the other worlds.) Given that, in W_3, it's my belief that CVR is true that leads me to utter U_3; it follows that it is that same belief that prompts me to utter U_1 and also U_2.

2.4 Burge's failure to distinguish literal meaning from cognitive content

U_3 does indeed have different truth-conditions from each of U_1 and U_2. But it doesn't follow that the thought underlying U_3 is different from the thought underlying U_2/U_3.

Not only does it not follow: it isn't true. What VR is telling me, and what I believe strictly on the basis of it, is CVR. It's true that I have other beliefs. I believe (incorrectly) that it's Max I'm seeing. But that belief is a derivative of my belief that CVR is true. And the belief that prompts me to utter U_3—the belief I'm expressing—is true, it being irrelevant who or what the plaid-shirt wearer turns out to be.

An extension of our story about W_3 makes this clear. In plain view of Sally and myself, this entity takes its face-plate off, revealing a complex web of circuitry, as opposed to various organic tissues. I now know that it's a robot, and, given my pre-existing knowledge that Max isn't a robot, that it

therefore isn't Max. Now that I know that it was a robot that I was seeing, am I to conclude that the belief embodied in U_3 is wrong? No. If I look at what is in fact a green car and say "that car is green," what I'm saying is correct and the belief embodied in that utterance is correct, even though I also happen to have the false belief that the car belongs to Smith when, in fact, it belongs to Jones.

Nor am I to conclude that VR wasn't veridical. VR was high-resolution and veridical. The problem had to do with the fact that, in having VR, the thing I was accurately seeing, though not identical with Max, looked just like him, and with the consequent fact that, on the basis of my correct belief in CVR, I erroneously assumed that it was Max I was seeing.

2.4.1 Burge's failure to distinguish literal meaning from cognitive content (continued)

To clarify and substantiate the points just made, let's modify what we said about the events taking place in W_2.

You and I are both looking at Gary. (We both have good vision.) I happen to know that the thing we're seeing really is Gary. I have pre-existing knowledge to the effect that nothing other than Gary looks quite like Gary.

> VG is my visual perception of Gary. From an experiential standpoint, VG is completely indistinguishable

from the perception that, in W_1, prompted me to utter U_1 and also from the perception that, in W_3, prompted me to utter U_3. VG is the perception that prompts me to utter U_2.

It follows that from a purely experiential standpoint VG is completely indistinguishable from VM and also from VR. It also follows that the belief embodied in U_2 is identical with the belief embodied in U_1 and also in U_3. U_1–U_3 all express the same belief. So CVR is the belief embodied in each of these three utterances.

Although I know that it's Gary I'm seeing, VG *itself* obviously doesn't tell me this. Nor does it tell me that, apart from Max, nothing in the universe looks just like Gary. These things were learned independently of VG. My

knowledge of them is *background* knowledge.

Thus, there is no individual x such that x is identical with Larry and such that *by itself* VG tells me that it's x specifically that I'm seeing. I do indeed *know* who I'm seeing. But it's VG *plus* my background knowledge that gives this information, not VG by itself. Given only what VG is telling me, I could be seeing Gary or Max or some third thing, and the correctness of what I believe strictly on the basis of VG doesn't depend on who (or what) it is that I'm seeing.

3.0 The gappiness of Burge thoughts

The erroneousness of Burge's position becomes painfully clear when we register its most immediate consequences. Of those consequences, here's the one I find most striking. Given two people who are psychologically identical, one of them may be having all sorts of thoughts while the other is a zombie—a mental blank who is no more having thoughts than a tomato.[176]

Remember that, in W_1, there is some individual x such that x = Max and such that U_1 (my utterance of "that guy is wearing a plaid shirt") is true exactly if:

(MP[177]) x is wearing a plaid-shirt.

This is true. Notice that it's a claim about language, about literal meaning—about what is literally meant by an *utterance.* It isn't, at least not directly, a claim about the content of my visual perceptions or my thoughts.

But Burge holds that, indeed, positions exactly similar to the one just stated about that utterance hold with respect to my perceptions and thoughts. Burge's holds that, in W_1, there is some individual x such that x = Max and such that the content of VM (the visual perception I'm having) is true exactly if:

(MP) x is wearing a plaid shirt.

Burge *also* holds that, in W_1, there is some individual x such that x = Max and such that the content of what I *believe* on the basis of my visual

perception is true exactly if MP.

This last claim has a very interesting corollary. Let W_4 be a world that, with one qualification, is like each of W_1–W_3. The qualification is that, at the time in question, Sally and I are hallucinating. In other words, I'm not really seeing a man who is wearing a plaid shirt; there's only empty space in the place where, so I believe, I "see" there to be such a person. But I *am* having a hallucination that is phenomenologically (experientially) just like VM (and, therefore, VR and VG). We'll refer to that hallucination as "VH" (short for "visual hallucination"). Prompted by that hallucination, I say:

(U_4) "That entity is wearing a plaid shirt."

If U_4 were correct, it would be because, in the place in question (the one at which I'm pointing), there was some individual w such that w was wearing a plaid shirt; and the proposition that U_4 would encode under that circumstance would be:

(bw) w is wearing a plaid shirt.

But U_4 is not correct. It's not true or false. It's abortive. If I point to a non-existent gnome on my bed, and I say:

(G_1) "that gnome weights 8 lbs,"

my utterance *presupposes* that there is such a gnome, and since that presupposition is false, my statement is neither true nor false. By contrast, if I pointed to the bed and said:

(G_2) "there is a gnome on my bed, and that gnome weights 8 lbs,"

that statement would be false. (G_1 and G_2 are supposed to be utterances of sentences, not sentences per se—sentence-tokens, not sentence-types.)[178] U_4 is abortive for the same reason that G_1 is abortive. It is neither true nor false. Since any expression that is true or false *ipso facto* encodes a proposition, and since any proposition is *ipso facto* true or false, it follows that U_4 fails to encode a proposition.

Burge rightly believes that U_4 fails to encode a proposition. But he believes—partly on the basis, it seems—that *VH* has no content. In other words, he believes that VH isn't giving me *any* message. What VH is giving me, Burge believes, is the following message-fragment:

(BLNK) __is wearing a plaid shirt.

The underlining (the ___) is meant to denote a gap or blank of some kind.

Burge also thinks that the thought I'm having on the basis of VH has a blank in it—that it's like BLNK. In Burge's view that thought (or non-thought, as the case may be) is *identical* with BLNK.

If Burge is right, what VH gives me has a hole in it, and therefore isn't really a message; and to the extent that it's based on VH, the mental activity that follows it doesn't mediate any *bona fide* thoughts. Since the mental activity underlying U_4 is obviously based on VH, Burge's view entails that no *thought* underlies that utterance.

Burge's position is preposterous. It's obvious that VH gives me a complete message. It tells me that there is an individual in a certain place who wearing a plaid shirt (etc.). More precisely, it tells me that:

(CVR) There is some entity x such that x has such and such characteristics, such that x is standing is in a certain place and such that x is wearing a plaid shirt.

CVR is a false message, of course. But that only proves our point, since false messages are complete messages. Burge's analysis entails that the message given to me by my visual experience is BLNK, not CVR. Since BLNK, unlike the message that is in fact given to me, is neither true nor false, Burge's analysis is false.

4.0 A corollary of Burge's externalism: the impossibility of knowing one's conscious thoughts any better than one knows facts about the spatially and temporally remote past

Let's continue to think about my situation in W_4. In that world, I *think* that

I'm seeing somebody—and I'm wrong. But there's something I'm not wrong about. In W_4, if, at the time I was having VH, I were asked "are you having a thought and, if so, what is it?" I'd say without any hesitation: "yes, I'm having a thought; I'm thinking that the fellow standing over there is wearing a plaid shirt." Surely my answer to the second question is right. I know what I'm thinking. What I'm thinking is wrong, since there's no man there. But I'm right to think that I'm thinking it.

But Burge's position entails the opposite. In the context in question, I'm not thinking *anything.* There's a blank where, had I been in W_1–W_3, I *would* have been having a thought. A consequence is that, when you ask me "are you having a thought and, if so, what is it?" I'd be *wrong* to say: "yes, I'm having a thought; I'm thinking that the fellow standing over there is wearing a plaid shirt." Why would I be wrong? Two reasons. First, I wouldn't be having a thought at all. I'd be having a would-be thought, a pseudo-thought, but not a thought. Second, since I wouldn't be having *any* thought, my belief as to what specifically I was thinking would also be wrong.

But that's obviously false.

4.1 The causal conception of conception

To appreciate the depth and severity of this problem, we have to make a few facts about Burge's views explicit. Here is Burge's view (this is paraphrase, not a quotation):

> (BV[179]) Let's take another look at the situation described a moment ago—the one in which, in W_1, you are thinking about Max. The reason you're thinking about Max, opposed to Gary or some other entity, is that you're on the receiving end of a certain causal process that began with Max. Your current thought about Max (i.e., your thinking that he's wearing a plaid shirt, etc.) resulted from a perception that you had of him. That perception was the result of changes to your physiological condition that were brought about by light-rays that bounced off of him. That is why the sense-perceptions in which those disturbances eventuated were perceptions, not of Gary, but of Max; and that is also why the thoughts in which those sense-

> perceptions eventuated were thoughts, not of Gary, but of Max. So your current thought—the one embodied in U_1—concerns Max, and no one else, because it "tracks back" to Max, meaning that it bears a causal relation of the just-described kind to Max.

In Burge's view, conception is to be understood in causal terms, not in descriptive terms. Why does he think this? Remember Putnam's story about Smith and Twin-Smith. Smith's mental contents are *descriptively* just like Twin-Smith's. So far as Smith has a mentally encoded description that picks out H_2O, it's "liquid that we bathe in, drink, is found in lakes [etc.]"; and so far as Twin-Smith has as a mentally encoded description that picks out XYZ, it coincides with Smith's . At the same time, when Smith says "water covers most of my planet's surface," he's making a statement about H_2O, not XYZ; and when Twin-Smith utters those same words he's making a statement about XYZ, not H_2O. Given that Smith's thoughts encode the same descriptions as Twin-Smith's thoughts, why do Smith's words concern H_2O whereas Twin-Smith's concern XYZ? The answer: Smith is on the receiving end of a causal chain beginning with H_2O, whereas Twin-Smith is on the receiving end of a causal chain beginning with XYZ.

The essence of conception, in Burge's view, isn't description, and is instead the causal back-tracking relation just described. Smith is thinking about H_2O because his thoughts track back to H_2O, not XYZ.

A story makes it clear how untenable this is. Right now it's 5:38 P.M., April 11, 2009. Let t be our abbreviation for that time. Right now, at t, I'm thinking about Descartes. More specifically, I'm thinking:

(D*) Descartes was a fine philosopher.

I must emphasize that I'm doing a very a thorough job of reflecting on what's going on in my head right now; and, on that basis, I can say with complete confidence that D* is among the things that I'm thinking. Next year, historical research makes it clear that Descartes never existed—that there was no Descartes. It was all a hoax, we find out; the documents—all forgeries. According to Burge, I'm thinking about Descartes only in so far as my thoughts track back to him in the relevant way. They can't do that if he didn't

exist. So when it's discovered that Descartes never did exist, it not only turns out that, at t, I wasn't thinking D*: it *also* turns out that D* doesn't even exist to be thought. Let us momentarily suppose that Descartes *did* exist. In that case, there is some x such that x = Descartes, and such that D* is correct iff x is a fine philosopher and—Burge concludes from this—such that, for that very reason, if x didn't exist, D* wouldn't exist either. Burge thinks that Descartes is an actual *constituent* of the proposition—of the *truth*—that he was a fine philosopher. (I myself don't think this. I discuss this in Chapters 3, 6 and 8, where I identify the fallacies concerning the nature of propositions that underlie this view of Burge's.)

So, supposing that Descartes turns out never to have existed, despite years of thought to the contrary, it will turn out, in Burge's view, that D* doesn't even exist to be thought. So I can't even think that I'm thinking it—*it* doesn't exist, after all.

But that seems very wrong. Right now, I'm thinking that I'm thinking that Descartes was a fine philosopher. If JMK* is an occupant of a world where Descartes never existed who is otherwise exactly like me, JMK*'s psychological condition at the current time is exactly like mine. But, whereas *I* am now thinking that I'm thinking that Descartes was a fine philosopher, JMK* is failing to have that thought. To the extent that what *I* am doing right now is thinking D*, JMK* is altogether failing to have a thought. He's shooting a cognitive blank. That doesn't mean that he isn't thinking *anything.* But *to the extent* that what I'm thinking right now is D*, he isn't thinking anything. That means that all of the psychological activity that, in my case, is mediating the thought that D* is, in JMK*'s case, failing to mediate anything. And supposing that, for a split second, D* were the one thought passing through my mind, it follows on Burge's view that JMK* is, during that split second, *not thinking anything at all.* Thus, a consequence of Burge's view is that, given two people who are *psychologically* identical, one of them may be having a thought whereas the other is thinking nothing at all. But not only is that clearly false: it's incoherent. Thought *is* psychological activity. Burge's view has the absurd result that thought isn't a psychological process.[180]

4.1.2 Burge's response[181]

Burge denies that his position is inconsistent with the presumption that people know what they are consciously thinking. To evaluate this claim of his, let's start out by saying what he *does* admit. Burge admits three things, which we'll refer to as "B_1," "B_2," and "B_3":

(B_1) No matter what mental states I have, it's a possibility that Descartes didn't exist. The documents on which my belief is based could have been forged (etc.).

Thus, even though (so I think, and so let us assume) I *do* know that Descartes existed (at least insofar as I know anything about what happened in France centuries ago), I don't know *for sure* that he did. Moreover, given only what my own mental states are, it's a theoretical possibility that he never did exist.

(B_2) For argument's sake, suppose that, in reality, Descartes really did exist. Let's also suppose that I'm now thinking:

(D*) Descartes was a fine philosopher.

A consequence of my (Burge's) position is that there existed some individual x, namely Descartes, such that D* is correct if and only if

(D_1*) x was a fine philosopher

and such that the mental events and structures involved in my having that thought are correct if and only if D_1* is correct.

(B_3) B_2 entails that *if* I were to be in the very same mental state I was in when I was thinking D*, *but* it happened to be the case that Descartes had never existed, then I would *not* be thinking D*. In other words D_2 entails that the very mental events and brain-events that *in fact* constituted my thinking D_1* would fail to constitute *any* thought. By virtue of having those mental and neural events, I'd be shooting a blank, not having a thought—I'd no more be having a *bona fide* thought than a tomato.

B_1–B_3 are all correct. B_1 is correct, since I can't know for sure that my

mental states correctly represent the external world or a fortiori that the inferences involved in my believing in Descartes' existence are correct—inferences that are made on the basis of possibly faulty perceptions of possibly faulty records concerning 17th century France. B_2 is correct, since Burge's position does indeed entail that Descartes must have existed if anyone is to think that he was a fine philosopher. And B_3 is correct since Burge's position does indeed entail that the very mental events involved in my thinking D_1* would fail to constitute my having any thought at all if, supposing that everything else were the same, Descartes had never existed.

B_1–B_3 entail that:

(B_4) If I'm now thinking that:

(D*) Descartes was a fine philosopher,

then I cannot be any more certain of the fact that I'm thinking D than I can be of the fact that Descartes existed.*

B_4 is wrong. I obviously know what I'm consciously thinking with greater certainty than I know about medieval France. I want to make it as clear as possible why B_1–B_3 entail B_4. Given B_2, I can think D* only to the extent that Descartes really did exist. Burge admits this. (He has to. B_2 is Burge's whole position.) Given B_1, I can know that he existed only to the extent that various beliefs of mine concerning the external world to be accurate—beliefs that presuppose the accuracy of many perceptions of mine and also of many historical documents. Burge admits this as well.

B_3 entails that there is no experiential difference—no difference that is even theoretically detectable—between how I am when I'm *actually* thinking D^*_1 and how I am in a world where Descartes never existed but is otherwise like the world where I'm actually thinking D^*_1.

B_3 entails that, if I am to know that I'm thinking D^*_1, or anything else about Descartes, I'd jolly well better know that he existed; and I'd therefore jolly well better know that various Descartes-related perceptions of mine

were accurate and, in addition, that various inferences of mine concerning various records of mine were also accurate. Which means that a *pre*condition for my thinking about Descartes is my knowing that Descartes existed Which, in its turn, means that, supposing that I'm consciously thinking about Descartes right now, I can't possibly know that fact with very much certainty at all. I can know it only with the certainty with which I know that Descartes existed.

But that's false.

Since knowing that Descartes existed obviously involves knowing a lot of other things about the external world—since it involves my knowing that I was in fact reading certain books at certain terms, that certain records in fact exist and were made under certain circumstances, etc.—my knowing a huge amount about the external world is a prerequisite to knowing that I'm thinking this or that about Descartes.

But that's false.

This reasoning is easily generalized to show that, if Burge's position is right, my knowledge that I am consciously thinking about anything x in the external world is no more certain than my knowledge that x existed and, therefore, is no more certain than my knowledge of the various things on which that knowledge is based.

To sum up, Burge's position is false, since it has the false consequence that my knowledge of my own conscious mental states is as tentative as my knowledge of occurrences in remote regions of time and space.

Around six years after he wrote "Individualism and the Mental," Burge became aware of this problem. He and various followers of his have spent the last 23 years attempting to show that, even though B_1–B_3 are all correct, Burge's position is completely compatible with the presumption that we know our conscious mental states better than we know about Descartes' existence.

To those who would join Burge in this heroic quest, I must say, quoting Sponge Bob Square Pants: "Good luck with that!"[182]

5.0 The consequences of Burge's big blunder: the

need to deny independently known truths about causation

But there are problems with Burge's position far worse than any thus far mentioned. To see what they are, let's revisit the situation involving Smith and Twin-Smith—the one where, according to Burge, one of them, but not the other, believes that he has arthritis.

By supposition, Smith and Twin-Smith are psychologically and physiologically exactly the same. Given this, it follows that the causal properties of Smith's thoughts coincide with those of Twin-Smith's thoughts. This means that the supposed difference in content doesn't make a difference, and is completely and totally inert. Although Smith's *belief* in the proposition expressed by A_3 (or whatever) obviously has effects (on what Smith does, etc.), and although the same is obviously true of Twin-Smith's corresponding belief, the alleged *difference* between these beliefs is absolutely without any consequences. It has no consequences for anybody's overt behavior or for anybody's internal, psychological condition.

5.1 Fodor's insight

Jerry Fodor (1987) was one of the first people to note this consequence of Burge's view. Fodor's argument is a brisk and incisive one. What is it, Fodor asks, for two numerically distinct things, x and y, to be causally different (i.e., to differ in respect of their causal properties)? It is for them to be such that, were they to switch places with each other but all other factors were left unchanged, the one would behave in the way in which, in actuality, the other behaves.

Here's the idea. Suppose that x and y are two billiard balls that are atom-for-atom duplicates of each other. At time t, x collides with a green billiard ball and causes that billiard ball to move. At the same time, y collides with a red billiard ball and causes that billiard ball to move.

It's clear that the *effects* of the one collision are different from those of the other. In the one case, the effect is to make a green ball move; in the other, it is to make a red ball move. But this doesn't mean that the two billiard balls differ in their causal properties. It means that a given causal

property has different effects in different contexts. Suppose that I'm moderately strong. In some contexts, that will have the effect that I easily lift the barbell I'm trying to bench press. (Such contexts include those where the barbell weighs under 100 lbs.) In other contexts, that will have the effect that I can barely make the barbell budge. (Such contexts include those where the barbell weighs over 350 lbs.).

But, assuming that my level of strength hasn't changed in the intervening time period, the difference between a situation where I'm trying to lift 50 lbs and one where I'm trying to lift 350 lbs isn't that my causal properties have changed; it's that one and the same set of causal properties are sufficient for certain jobs but insufficient for others. You can't make a rocket out of plywood. But you can make a birdhouse out of plywood. That doesn't mean that the causal properties of plywood vary, depending on whether they're composing a rocket or a birdhouse. It means that those causal properties, though sufficient for certain purposes (e.g., bird-house building) are insufficient for others (e.g., rocket building).

In his landmark work, *A system of Logic,* John Stuart Mill made it clear that the just-described criterion of causal identity is presupposed by all experimentation and, setting aside the empirical fact that scientists rely on it, is correct as a matter of logic.

This is not to detract from Fodor's achievement. Fodor stated that criterion more sharply and clearly than any author that I know of who preceded him. Plus, Fodor deserves credit for seeing the bearing of this principle on Burge's views. Fodor deserves a lot of credit for many positive developments in recent analytic philosophy, as we'll see.

5.1.2 Why thoughts must have causal powers

Causality is indistinguishable from spatiotemporal existence. It makes no sense to say that x occupies a certain region of space-time but has no causal properties. This is because occupancy is itself a causal notion. My occupying the chair I'm in is indistinguishable from my altering the conditions that other things (e.g., you) must satisfy if they are to occupy the chair. If I weren't in the chair, you wouldn't have to displace me to occupy the chair; and given that I *am* occupying the chair, you *do* have to displace me to occupy it yourself. Which, though certainly possible, would require an otherwise

unnecessary expenditure of energy on your part.

Given that Smith and Twin-Smith are absolutely the same in respect of their causal properties, it follows that the alleged differences between the one person's mental content and the other's is causally impotent. A moment ago we saw that it makes no sense to posit the existence of anything spatiotemporal that has no causal powers at all. The just-cited alleged difference has no causal powers. Therefore, it makes no sense to suppose that it exists. Therefore, Burge is wrong.

5.2 Burge's response to Fodor

Burge's response to Fodor is rather curious. Burge *denies* that billiard balls x and y are causally identical. Burge says that the causal properties of an object are to be understood in terms of what it *actually* does and not at all, contrary to what Fodor says, in terms of what it *would* do in different circumstances.

A story will help make it clear what Burge is saying. On Monday, I run a 6-minute mile without wearing a 20-pound backpack. On Tuesday, I run a 7-minute mile wearing such a backpack. On both days, I run on the same path. The weather is the same. I'm no worse and no better rested or nourished on the one day than on the other.

According to Burge, my level of fitness has declined, the reason being that I ran the mile more slowly on Tuesday than on Monday. It's irrelevant, says Burge, that, on Tuesday, I *would* have run the mile as quickly as I did on Monday *if* I hadn't been wearing the backpack. Why is it irrelevant? Because, says Burge, the causal properties of physical objects are to be understood in terms of the peculiarities of the environments in which they happen to be embedded.

Let's look at the example that Burge himself uses. Let A and B be two very large celestial bodies that are atom-for-atom duplicates of each other. A is revolving around a star. (A is a planet.) B is not revolving around any star. (So B is not a planet.) Burge says (rightly) that the concept of a planet is an explanatorily important one. We explain at least some things in terms of the fact that some things are planets. Burge concludes from this that A and B have different causal properties. It is not, says Burge, that A and B merely differ in their behavior. It is that, by virtue of being a *planet*—albeit one that

is exactly like B down to the last atom—A has causal virtues that B lacks.

Is Burge's argument a good one? No. Suppose that Smith and Jones are atom-for-atom duplicates of each other, but that Smith has a nephew whereas Jones does not. Obviously there are, or at least could be, behavioral differences between Smith and Jones that are to be explained in terms of the fact that the one, but not the other, is an uncle. Why did Smith spend an hour shopping for toys? Because he's six-year old Timmy's uncle, and Timmy loves toys, as Smith knows. Why didn't Jones spend an hour shopping for toys? Is it because Jones is mean and stingy, whereas Smith is kind and generous? No. It's because Jones, unlike Smith, isn't an uncle and thus does not (at least not for that reason) have go toy shopping. Burge's position concerning planets, if transposed to this other context, is that Jones is in fact a very different person from Smith. Smith is a nice guy. (Let's assume for argument's sake that his toy shopping is an expression of genuine love for his nephew.) But Jones, not having gone to the trouble of buying toys for his non-existent nephew, is not such a nice guy, even though Jones is just as likely as Smith to buy a present (or otherwise behave amicably towards) the people in his life who, unlike his nephew, actually exist.

Because A is embedded in a certain way in a certain solar system, it's going to do things that B won't do. It'll move in ways that B will not. It'll draw the astronauts who land on A towards itself, whereas B will not do that, since (we may suppose) no astronauts bothered to go to B.

Do any of these facts entail that A has different causal properties from B? No. In describing A as a "planet," one is saying that it is embedded in a certain way in a certain environment. In describing B as a non-planet, one is saying that it is not embedded in that way in that sort of environment. Burge is quite right that what things do depends on the circumstances. If I'm underwater, I don't breathe; if I'm not, I do. Since A is part of a solar system, it does certain things; since B isn't, it doesn't. But this doesn't show that A and B differ in respect of their causal properties from each other. It shows that:

(i) A's being embedded in a certain environment in a certain way

has different effects from

(ii) B's being embedded in a totally different way in a completely different sort of environment.

(i) is not identical with A. (ii) is not identical with B. (I am sitting in a chair right now. But I'm not identical with *the state of affairs consisting of JMK's sitting in a certain chair at 1:22 P.M.* on April 9, 2009.) A's being a planet is identical with (i)'s being the case. Since A's being a planet is no more identical with A than I am identical with the fact that, at this instant, I'm sitting in a certain chair, and since the same thing *mutatis mutandis* holds of B, it would be deeply illogical to accept Burge's judgment that A and B differ in their causal properties. In order for that judgment to be correct, A *would* have to be identical with the fact that it was embedded in a certain way in a certain sort of environment, and by parallel reasoning you *would* be identical with the fact that, right now, you are holding a certain book in your hands. But so far as it isn't obviously false, this position borders on meaninglessness.

Despite all these problems with content-externalism, and despite some comparably gruesome ones that I haven't even mentioned[183], content-externalism still has some loyal followers who are indefatigable in their attempts to make content-externalism work. To these tenacious individuals, I can only say, quoting the great Marion Barry: "Let it go!"

PART III

Knowledge and Inductive Inference

Chapter 10

Knowledge

1.0 The ambiguity of the word “know”

The term “know” is ambiguous. Sometimes it refers to awareness of *objects,* as in “I know Smith.” But oftentimes it refers to awareness of *truths,* as in “I know that snow is white.”

Other languages use different verbs to mark this difference. For example, in Spanish, when one speaks of knowing *people,* one uses the verb “conocer,” and when one speaks of knowing truths, one uses the verb “saber.” German also uses different words to mark this distinction (“kennen,” for knowing people, and “wissen” for knowing truths). Russian does not. It is like English, and uses one verb—“znat”—for awareness of both people and truths.

This last fact is suggestive. If English were the *only* language that had just one word for both kinds of awareness, its doing so would be a mere idiosyncrasy—a random fact about one language that didn’t reveal anything inherent in human thought. But the fact that *other* languages do the same thing suggests that there is a general tendency *not* to distinguish these two very different kinds of knowledge. We will find that, in fact, this tendency of ours has a solid epistemological basis and is therefore not a logical *faux pas.* However counterintuitive it may seem, knowledge of people *is*, we will find, knowledge of truths.

1.1 The philosophical meaning of the word “know”

In philosophy, when we talk about “knowledge,” we are primarily, but not exclusively, concerned with knowledge of *truths.* Nonetheless, we will find that the distinction between knowledge of truths, on the one hand, and knowledge of individuals, on the other, is of great importance.

We will find later that the word “know” is actually ambiguous in additional ways. For example, psychologists and philosophers often talk about “procedural” knowledge. Any case of anyone’s being able to do anything is a case of procedural knowledge. My ability to play the piano is a

case of procedural knowledge; so is my ability to speak English; and so is my ability to tie my shoes.

Many deny that procedural knowledge is *bona fide* knowledge. This is the exact opposite of the truth. What the virtuoso knows procedurally, the incompetent neophyte knows only discursively. As it deepens, knowledge becomes increasingly proceduralized. Non-procedural knowledge is poorly integrated knowledge; it is knowledge that hasn't yet established the links to other pieces of knowledge that are needed to mobilize it. Procedural knowledge is knowledge that is sufficiently well-integrated to have established these links. It is therefore indefensible to hold that procedural knowledge is non-knowledge. We'll say more about this in Section 9.3.

We'll find that the word "know" has meanings *additional* to those already discussed. But for the time being, when the word "know" is used, assume that it is referring to knowledge of truths, unless there is an explicit indication to the contrary.

It should be pointed out that, even though, for the reasons given, the word "know" *seems* to have many meanings, some philosophers maintain that ultimately all knowledge is identical with knowledge of truths—in other words, that awareness of objects and also procedural knowledge, along with the other kinds of knowledge that we will talk about, are ultimately just different ways of knowing truths. But this is extremely controversial.

2.0 Knowledge as justified true belief

One cannot know something false. Of course, people often firmly *believe* that they know things that are in fact false. People can believe that the Earth is flat with as much conviction as they can believe that the Earth is round. The distinction between a true belief and a false one is not always marked by internal psychological or emotional differences. For example, one cannot tell whether a belief that one has is true or false merely on the basis of the intensity of that belief. In any case, it is a simple fact that only truths can be known.

All cases of knowledge are cases of belief. If you know something, you believe it. If you know that Paris is the Capital of France, you believe it. If, speaking non-ironically, somebody said "I know that Paris is the capital of

France, but I don't believe it," that person would be uttering nonsense.

But not all instances of belief are instances of knowledge. We've already seen one reason for this. False beliefs are not knowledge. But there is another reason why not all belief is knowledge. A belief must be *justified* if it is to be knowledge.

A lucky guess is not knowledge. Suppose that you're taking a multiple-choice test and you come across the following question:

With what digit does Professor Kuczynski's phone number end?

(a)
9

(b)
3

(c)
2

(d)
0

(e) none of the above

A lucky guess will be enough for credit on the test, but it won't be enough for knowledge. You have must have some good reason for your choice. Your opinion as to what the right answer is must be *justified* for it to qualify as knowledge.

There are a lot of ways that you can have such a justification. Maybe I've told you my number. Maybe you saw it on one of my class handouts. Maybe you hired a private detective to find it out. All of these would be good sources of justification. But without some such source, you don't have knowledge.

Also, spurious reasons that happen to lead to the right answer are not enough for knowledge. Suppose that you believe, wrongly, that all people with red hair have cancer; and also suppose that, for that reason, you believe that your red-haired friend Bob has cancer. It may be that Bob *does* have cancer. But your reason for believing this is spurious; so you don't *know* it.

It is a general principle of logic that false premises can lead to true

conclusion; misinformation can lead to correct beliefs. Of course, accurate information is much more *reliable* than misinformation. A person whose methods are sound and whose premises are correct is considerably more likely than he would otherwise be to find out the truth. But false premises and unsound methodologies can lead to truth. When it happens, it's a kind of accident, as our previous example shows. But accidents do happen.

So for something to be knowledge it must have the following three characteristics:

(1) it must be true;

(2) it must be a belief; and

(3) it must be justified.

Thus, all cases of knowledge are cases of justified true belief.

2.1 What constitutes justification?

We must define a few more terms before we can answer this question. The term "perceptual" means having to do with sense-perception—that is, with sight, hearing, sound, taste, smell, bodily sensations (e.g., stomach aches), and also sensations of motion and resistance. I know that there is a computer in front of me because I can see it. (In any case, this is part of the story: there's more to it, as we'll see later.) I know that the stove is hot because I touched it. I know that one cup is heavier than another because I feel a greater degree of resistance when I try to move the one than when I try to move the other. (Notice that this last kind of perception isn't identical with one's sense of touch, even though it involves the use of touch.)

The term "introspective" refers to the direct awareness that one often has of one's own mental contents. Right now I am aware of the fact that I am having thoughts—that, more specifically, I am having thoughts about the theory of knowledge (*I'm writing about it*), about this chapter (*I hope that I've finished writing it by tomorrow*), and about tonight's dinner (*it'll be chicken*). My knowledge of this is not based on the use of my senses: I don't *see* or *touch* those thoughts. At the same time, that knowledge seems not to involve my making inferences: I seem to be directly, non-inferentially aware that I am having them.

Bearing all of this in mind, let us consider the question: under what circumstances is one justified in having a belief? Many different kinds of facts can be known; and, as we'll see, the way in which a belief must be justified for it to be knowledge depends on the category into which the fact in question falls. But there are ultimately only three sources of justification: (a) perceptual evidence, (b) introspective evidence and (c) rational inference. Some cases of knowledge involve *only* (c); some cases of knowledge are a combination of (a) and (b); the majority of cases involve a combination of both (a) and (c). Some philosophers[184] have argued that whenever (a) is involved, so is (b).

2.2 Can testimony constitute a source of justification?

There is much that a given person knows that he hasn't personally verified—much that he knows because he is *told* it. Such telling can take various forms: it can consist of written or spoken speech, photographs, and drawings. (This list is not complete.)

I haven't personally looked at the Earth from outer space. But I still *know* that it is approximately spherical. How do I know this? Because I've been told it. Using speech, writing, photographs, drawings, and possibly other media, various people have told me various things that justify me in believing that the Earth is spherical. For similar reasons, I have at least some knowledge about what kinds of activities are likely to bring about lung cancer and about the political situation in distant parts of the Earth, even though I have not personally done cancer research or personally visited those parts of the Earth.

But, of course, not *all* beliefs that are based on testimony qualify as knowledge. First of all, much of what we are told is false. But even true beliefs that are based on testimony are not always knowledge. To qualify as knowledge, testimony-based belief must (a) come from a reliable source and (b) the person being given the testimony must have independent evidence of the reliability of that source.

A story may clarify this. I belong to a cult. My cult leader wants me to

drink a certain beverage. He has told me that it is extremely nutritious. As it happens, it *is* nutritious—the cult leader has told me the truth. Further, I *believe* that the beverage is nutritious. But the reason for this is *not* that I have any actual evidence that the cult leader is well-informed or truthful; the reason is that, for purely emotional reasons, I believe everything the cult leader tells me.

Here I have a true belief—I believe, correctly, that the beverage is nutritious. But do I *know* this fact? No. And the reason is that I don't have good reason to believe that the cult leader is in fact telling the truth—even though, in this case, he is telling the truth. The data that I have concerning the cult leader—more specifically, concerning the likelihood of his stating truths as opposed to falsehoods—does not warrant my accepting what he is saying.

Here's another story. A world-renowned physician has told me that I have arthritis. His opinion is based on his expertise and also on the results of various tests that he has performed on me. He is right—I do in fact have a mild case of arthritis. Further, I am aware of the fact that he is an arthritis expert and that he has conducted various tests.

In this case, I obviously *do* have reason to believe the testimony I am being given. My belief is therefore justified and it is, or at least could be, a case of knowledge. (I say "or at least could be" because, for reasons that we'll discuss later, even beliefs that are true and justified aren't always knowledge.)

So a testimony-based, true belief can be knowledge. But it is necessary that the person who has that belief have some reason to believe that the source of the testimony is accurate.

Notice that testimony-based belief involves a combination of perceptual knowledge and also of rational inference. You must perceive the words that are being told to you. (You hear them or see them, depending on whether they are written or spoken.) You must also have perceived some kind of evidence of the truthfulness of the source of testimony. Finally, from those two sources of perceptual information, you must infer that the source of testimony is reliable. So testimony-based knowledge involves a combination of (a) and (c).

2.2.1 A point concerning testimonial knowledge

When we're growing up, we're told many true things by parents, teachers, and other figures of authority. During our earlier years, we obviously don't have independent corroboration for the testimony of these people. And yet we're inclined to say that we *know* the things they told us.

What I believe to be going on is that, at first, we *didn't* know these things. We just believed them. But then we either came to have evidence of the reliableness of the people who told them to us or we discovered them for ourselves.

3.0 Knowledge as hyper-justified true belief

Not all cases of justified true belief are cases of knowledge. It is only when one has a belief that is true and that is justified in a rather specific way that one has knowledge. More specifically, one has knowledge only when one has a belief that is true and justified and, moreover, *one's being justified in holding that belief is non-accidentally related to that belief's being true.*

A story will help clarify this. You work at some company. Smith is a fellow employee of yours. He is an excellent worker; the boss likes him; he's overdue for a promotion; and the boss, who is honest and rewards good work, has just announced that he's going to give a promotion to somebody (but he hasn't said who). Further, you know all of these facts. You also know that Smith drives a Ford. On the basis of this information, you form the belief that:

(*) somebody who drives a Ford is about to be given a promotion.

Obviously this belief is justified, since your reasons for holding it are solid.

And that belief turns out to be true—but not for the reason that you have in mind. Smith does *not* get the promotion. The boss gives the promotion to Jones. Jones is a terrible employee, as you know. But what you *don't* know, and had no way of knowing is that the boss owes a huge favor to Jones' dad, who is a local and exceptionally vindictive Mafia boss. So, acting with an uncharacteristic disregard for merit, the boss gives the promotion to Jones. By sheer coincidence, Jones drives a Ford.

In light of this, let's look at your belief in *. First of all, that belief is true —a Ford driver (namely, Jones) *was* given a promotion. Second, it's justified.

We've already seen why. You were justified in believing that Smith, whom you knew to be a Ford driver, was going to be a given a promotion. That belief turned out to be wrong—but it was still justified.

So your belief in * is true and justified. It is thus a justified true belief.

But is it knowledge? No. It's pure coincidence that Jones, who you had no way of knowing would get the promotion, drives a Ford.[185] You're right a Ford driver would get the promotion, and your belief is justified. But you still believe it for the wrong reasons. (A justified belief can thus be had for the wrong reasons.) And that's why it isn't knowledge.

This shows that there can be justified true beliefs that are not knowledge. It shows that, for something to be knowledge, it is not *sufficient* that it satisfy conditions (1)–(3), even though, for the reasons discussed earlier, it is *necessary*.

The argument just stated was given by Edmund Gettier (1963). Bertrand Russell (1912) gave a similar argument. You look at a clock that is in fact broken but that you believe, and have every reason to believe, is in good working order. According to the clock, the time is 4:00 P.M. In fact, that *is* the actual time. But the clock is broken, and it is therefore a fluke that it gave you the right time. Your belief that it is 4:00 P.M. is thus true and justified, but it isn't knowledge.

3.1 Knowledge as non-accidentally justified true belief

We've seen that for something to be knowledge, it isn't enough that it be a justified, true belief, and that it must therefore satisfy some fourth condition. There is considerable debate as to the precise nature of that fourth condition, but it is seems clear that, when stated very approximately, it must be this:

(4) if a given belief is to be knowledge, there must be a non-accidental connection between one's reason for having that belief, on the one hand, and the fact in virtue of which that belief is true, on the other: justification and truth must be non-accidentally connected.

So for something to be knowledge is for it to satisfy the following *four*

conditions:

(1) it must be true;

(2) it must be a belief;

(3) it must be justified; and

(4) there must be a *non-accidental* connection between the justification for the belief and the fact which makes the belief true.

Thus:

(K) knowledge is a *non-accidentally* justified true belief.

There must, in other words, be some kind of *principled,* as opposed to a strictly fortuitous, relationship between the justification of the belief, on the one hand, and its being true, on the other.

But K is much too vague to be considered a solution to the problem at hand. What does the word "non-accidentally" mean in this context? Until we know the answer to that, we don't know what knowledge is. We'll now try to supply that answer.

3.2 This position refined

Let's consider Russell's clock story. You were right to believe that it's 4:00 P.M., and you are also justified in having that belief. Your immediate justification consists in your belief that:

(RT) The clock is now telling me that it's 4:00 P.M.

But *by itself* a belief in RT wouldn't be enough to justify you in believing that it's 4:00 P.M. If you had no grounds for accepting RT, or only bad grounds, your acceptance of it would not provide you with any justification for believing the time to be 4:00 P.M. But you are justified in accepting RT; and here is that justification:

(OB) In my extensive past experience, the clock has always been right. I also know that the building superintendent, who is in charge of making sure all the clocks are working, is very capable and diligent. These facts ensured that, if the clock says that it's a certain time of day, that's because

it really is that time of day. That's why *the clock is now telling me that t it's 4:00 P.M.*

Of course, the italicized part of OB is identical with rt.

Notice that OB is an explanation of a fact—of the fact that the clock is now telling you that it's 4:00 P.M. It is on the basis of your acceptance of RT that you believe the time to be 4:00 P.M. But OB is the wrong explanation of RT. The right explanation is that the clock is broken and, for that reason, always reads 4:00 P.M.

A story will help us move forward. You believe that Larry is in Texas, and your belief is justified. Earlier today, you saw Larry get on a plane. By all indications, the plane was Texas-bound. (It's flight 71 and, according to the overhead monitor, that flight is bound for Texas.) When Larry got to Texas, he called you on his cell phone/video phone. He was standing next to a well-known Texas landmark (or, at the very least, a perfect duplicate therefore). On that basis, you believe that, indeed, Larry is in Texas. And, in fact, he *is* in Texas.

Moreover, no off-the-wall, Gettier-type event infiltrated this sequence. In other words, what happened is nothing like the following:

> Alternative Scenario (AS): Larry got on the plane. But Martians re-routed the plane to Barbados. For some reason, they then transported Larry to Texas. In Texas, somebody other than Larry, who looks and sounds just like him, calls you and tells you that "he" (which you take to mean "Larry") safely arrived in Texas. On that basis, you believe that Larry is in Texas. Which he is, and which you're justified in believing, but which, clearly, you don't know to be the case.

No—neither AS nor anything at all like it occurred. What happened is exactly what you thought happened. And you therefore know that Larry is in Texas. (Let's say that this story—the one where you *know* that Larry is in Texas—takes place in the "real world," even though, of course, it's no less a hypothetical than AS.)

Why do you know it in the real world, but not in the world described by AS? Let's look at the data on which, in the real world, your belief is based. Although you see Larry get on the plane, you don't see him deplane in Texas.

You *do* see various images (on your cell phone/video phone) that are consistent with his being in Texas; and you encounter a lot of other things that, were Larry not to be in Texas, would be very hard to explain. Your belief that Larry is in Texas (after sitting on a certain plane, etc.) is not only justified: it also constitutes a correct explanation of the data that justifies your belief that he is in Texas. Your belief that, after sitting on a certain plane (etc.), Larry is in Texas correctly explains *why* your cell phone is giving rise to these various images and noises.

Thus, your belief is not only justified, but *hyper*-justified. A belief is hyper-justified if it's justified and, moreover, embodies a correct explanation of the data that is doing the justifying.

In AS, your belief is justified, but not hyper-justified. Your belief is based on the same images and, in general, the same data that it's based on in the real world. But, in AS, your beliefs as to what caused those images and noises to occur are wrong. (You explained their occurrence by wrongly supposing that Larry produced them after being transported by non-Martians to Texas.) So, in AS, your justified belief isn't knowledge because it isn't hyper-justified. I propose, then, that:

(KN) Knowledge is hyper-justified true belief.

4.0 Conceptual (analytic) versus empirical (synthetic) knowledge

Some knowledge is purely conceptual; and some knowledge has a non-conceptual component. Knowledge that is purely conceptual is described as "analytic." Knowledge that has a non-conceptual component is described as "synthetic."

Consider the statement:

(#) Circles are figures of uniform curvature.

is true. And you know that it's true. You thus have knowledge of #. How does one become justified in believing #? Answer: by examining the statement itself. More specifically, by examining the components of that statement—the concepts *circle, curvature, uniform,* and so on. So your knowledge of # is purely conceptual. In other words, it is "analytic."

Analytic knowledge is so-called because it is acquired by *analyzing* the object of belief. Whenever statement-analysis is sufficient, by itself, to yield knowledge, that knowledge is analytic.

Most knowledge is non-analytic. Consider the statement:

(##) Donald Rumsfeld was born in the United States.

An examination of that statement is not enough to know whether it is true or false. It doesn't matter how much you dissect it and examine its components. In order to be justified in having a view, one way or the other, as to whether it is true, and in order therefore to *know* whether it is true, it is necessary to consider matters of fact of which no amount of conceptual analysis will apprise you. It is necessary to look at birth records or to ask informed people, or some such.

Thus, if one knows that ## is true, that knowledge is *non*-conceptual. In other words, it is *non*-analytic or, as philosophers sometimes say, *synthetic.*

4.1 More on the analytic-synthetic distinction

Although some philosophers have held otherwise, I think it is fair to say that synthetic knowledge is identical with knowledge of spatiotemporal facts, whereas analytic knowledge is knowledge of properties.[186]

To know ##, for example, is to know something about the spatiotemporal world—the world of things that occupy space and time. By contrast, to know # is to know about unchanging, non-temporal (and therefore non-spatial) relations among properties.

In this chapter, we will operate on the reasonable, though debatable, assumption that all synthetic knowledge is knowledge of some fact or other about the spatiotemporal realm. So, in this context, the terms "spatiotemporal knowledge" and "synthetic knowledge" may be used interchangeably. In Chapter 18, it is shown that, in fact, synthetic knowledge is identical with spatiotemporal knowledge.

5.0 Knowledge of specific facts versus knowledge of dependence-relations among facts

There is an important division within the category of spatiotemporal, or synthetic, knowledge. There is knowledge of *specific* facts, and there is also knowledge of *causal* or *nomic* relations among facts.

The term "nomic" means "having to do with natural law." Consider the fact that there is a combustion whenever a lit match is brought in contact with gasoline. Obviously that is not a coincidence. There is clearly some kind of principled relationship between, on the one hand, something's being gasoline that is brought into contact with flame and, on the other hand, something's combusting. Some kind of *causal* relationship is at work.

Technically speaking, not all nomic relations are causal. But the distinction between causal and nomic relationships is a rather subtle one, and in this chapter we will ignore it. (In Chapter 17, Sections 9.0–9.6, it is made clear why not all nomic dependencies are causal dependencies.)

Causal relations are instances of *natural laws.* If x causes y, that is because there is some law of nature to the effect that things that, in some relevant respect, resemble x always bring about things that, in some relevant respect, are like y.

Causal relations typically involve *sequences* of events. But not all sequences of events are causal. When I bring my hands together quickly, a noise occurs—there is a clapping sound. That sequence of events is clearly causal in nature: the collision of my hands caused the sound to occur. But suppose that, just after I bring my hands together, one of you suddenly feels a tickle. That is not a causal sequence: it's just a coincidence.

The natural law of which a given causal sequence is an instance is known as the *covering law* for that sequence.

To know a natural law is to know a fact about the spatiotemporal world. Natural laws are not in the same category as #: such laws are not of a strictly logical, philosophical, or mathematical nature. It is not possible to know of natural laws merely on the basis of conceptual analysis. Such knowledge can be acquired only through investigation of the spatiotemporal world. It is only through such investigation that one can know that the speed of light is 186,000 mph/second, that objects gain mass as they accelerate, that the force with which two bodies attract each other is proportional to their masses.

Of course, some natural laws strike us as extremely obvious. We find it

obvious that unsupported objects fall. But that is not because such knowledge is analytic; it is only because the relevant spatiotemporal knowledge constantly forces itself upon us.

Even though knowledge of natural law is spatiotemporal knowledge, such knowledge tends to have a highly *general* character and is thus, in that respect, comparable to many cases of analytic knowledge (e.g., one's knowledge of #). Knowledge of the laws of physics is very general knowledge. To know such laws is to know this or that specific fact, and is instead to know a fact about how events *in general* are interrelated.

But most spatiotemporal knowledge is highly specific. You know where you live; you know what toothpaste tastes like; you know where you parked your car; you know that Barack Obama is now the president. Such knowledge concerns matters of specific fact and not dependence-relations holding among facts in general.

It must be emphasized, however, that even though knowledge of natural law is general and is, from that viewpoint, similar to knowledge of logical or philosophical principles, knowledge of natural law is synthetic, whereas knowledge of logico-philosophical truths is analytic.

6.0 Knowledge of the future, the past, and the merely possible

Many future facts are not known and many future facts cannot *possibly* be known. But surely we are able to know at least *some* facts about the future.

One can know about the future because (i) on the basis of sense-perception, one can know about the present and (ii) one can know general principles, or causal laws, that link the present with the future. Right now I know that there is water in the pitcher in front of me. I also know of the general fact that, under conditions relevantly like those that obtain in my kitchen, an open, overturned pitcher containing non-viscous liquid will release that liquid. Further, I know that I have an intention to tip the pitcher, so as to fill the glass of water in front of me. On the basis of these pieces of knowledge, I can obviously have a non-accidentally justified true belief to the effect that, in a moment, water will be flowing from the pitcher into the glass.

Knowledge of the future always has two ingredients: first, knowledge of

present conditions and, second, knowledge of causal laws that, given circumstances similar to those known to obtain, yield certain outcomes. In some cases, the causal laws in question are of a relatively non-theoretical character. This was the case in the example, given a moment ago, concerning the water in the pitcher. In other cases, knowledge of the future may have an extremely theoretical character. This would be the case when, on the basis of relativity theory, a physicist predicts that the mass of a certain body will increase by a certain amount, by a certain time. But, in all cases, knowledge of the future is constituted by knowledge of causal laws[187] that link known current states of affairs to future states of affairs.

6.1 Knowledge of the past

We can obviously have knowledge of the past. There are two forms that such knowledge takes. In some cases, knowledge of the past is *direct:* one actually *remembers* the fact in question. I remember driving to work yesterday. My knowledge of this fact doesn't involve inferences made on the basis of causal laws.

But in other cases one's knowledge of the past *is* indirect, and all such knowledge *does* involve some knowledge of some kind of causal law. We know facts about the political institutions of ancient Rome. We obviously don't remember such facts, since we didn't personally experience them. So how do we know them? We have knowledge of certain current facts (e.g., certain currently existing records and ruins); and we *also* have knowledge of causal principles on the basis of which, given our knowledge of those current facts, inferences concerning the past—specifically, concerning what went on in Rome thousands of years ago—can be made.

Here's a simple example of the principle in question. You know that if you press an ink-filled pen against paper, inky lines will be deposited on the paper. (You also know, so let us suppose, that such lines are not likely to be deposited on paper in any other way.) This is knowledge of a causal principle. On the basis of your knowledge of this principle, plus your knowledge of current conditions, you can acquire *either* knowledge of the future *or* knowledge of the past. If you see that Smith is about to press his pen against the paper in his notebook, you know a fact about the future—you

know that inky lines will soon result. On the other hand, if you see inky lines that were clearly made by a person, you know a fact about the past—you know that somebody dragged a pen across a certain piece of paper. Our knowledge of past events that we don't personally remember (e.g., facts about ancient Rome) falls into this second category.

When it is indirect, knowledge of the past has a structure that is similar to knowledge of the future. More specifically, past knowledge can be thought of as *inverted* future knowledge. We know the future when, using known causal laws, we project forward from known current conditions. We know the past when, using known causal laws, we project backwards from such conditions.

6.2 Counterfactual knowledge

We can have knowledge of what *would have been.* Even though Smith did not in fact fall off the ledge, I know that:

(S) *if* Smith had fallen off the ledge, he would now be dead or severely injured.

Propositions like S are called *counterfactuals* . In general, a counterfactual statement is one of the form *if P, then Q,* where it is assumed that P is false.

Knowledge of counterfactuals is identical with knowledge of relations of either logical or causal dependence. To know S is to know of some causal law that, given a case of somebody's falling from a great height, leads to that person's breaking his or her leg. To know that "if x were a square, x would have four sides" is to know that "x has four sides" is a logical consequence of "x is a square."

Many contemporary philosophers take counterfactual statements to be statements about events in *other universes.*[188] So they take:

(K) "If Kennedy hadn't been assassinated on Nov. 22, 1963, he would have been reelected in 1964"

to mean that:

(K#) There is some other universe where Kennedy was *not* assassinated on Nov. 22, 1963, but, up until that time, was as much like our world as that fact

allows it to be. In that universe Kennedy was reelected in 1964.

And they argue that:

(S) “If Smith, who is 6-feet tall, were half his current height, he’d been 3-feet tall”

means that:

(S#) There is some alternative universe where Smith is half his current height. In that universe, Smith is 3-feet tall.

This is preposterous. Any knowledge that we have about other universes is parasitic on our knowledge of modal truths. First, we know the counterfactuals. Only then do we know what’s going on in other worlds. It’s only thanks to our knowledge that 6 divided by 2 is 3, along with out knowledge that Smith is six feet tall, that we know that Smith would be 3 feet tall in a world where he were half his actual height. It’s only thanks to our knowledge that Kennedy was an enormously popular first-term President, along with our knowledge that popular incumbent Presidents win re-election, that we know that Kennedy would have won in 1964 had his life not been cut short.

These pedestrian points show that other worlds are useless in the way of giving us modal knowledge. It also shows that modal truths don’t hold in virtue of what goes on in other worlds. S is obviously correct. Whatever it is that makes it correct, we don’t have to leave this world to have access to it. It follows that, given any true counterfactual statement S, it is in virtue of what happens in this world, and this world alone, that S is true.

The real meanings of S and K are:

(S*) the statement that Smith is 6-feet tall entails that the statement, for all values of x, ‹x is half Smith’s height› entails ‹x is 3-feet tall.›

and

(K*) “Kennedy’s assassination caused an otherwise unthreatened Kennedy not to be reelected.”

For further discussion of the nature of counterfactual truth, see Chapter 17.

7.0 Theoretical versus observational knowledge

It seems that much of what we can know comes *directly* from sense-perception. I open my eyes in a well-lit room, and I can just *see* that there are chairs and people there.

But much knowledge is theoretical. We don't have direct knowledge of electrons or even of gravitational fields. We have direct knowledge of the *effects* of these things. (More accurately, we have *relatively* direct knowledge of such effects. We'll see in a moment that our knowledge of these effects is itself indirect, even though it's obviously less so than our knowledge of electrons, etc.) But in characterizing something directly perceived as the effect of "gravitation" or of the activity of "electrons," we are going well beyond what we can literally see (or hear or touch, etc.), and we are assuming the existence of things whose existence can be justified only by *theories* of some kind or other. A theory is an explanation of what is directly experienced in terms of what is not directly experienced and whose existence is therefore not directly known.

Freud and others have posited the existence of mental activity that we are not directly aware of and of which we are therefore "unconscious." One obviously can't be directly aware of unconscious mental activity. (Any mental activity that one is directly aware of is *ipso facto* conscious.) Thus, any knowledge that we have of the unconscious is inferential. In addition, such knowledge is also *theoretical.* But here we must take a moment to clarify the distinction between inferential and theoretical knowledge.

A theory is an integrated set of statements that, if correct, describe some sector of reality that cannot be known strictly through sense-perception or introspection. All theoretical knowledge is inferential, but not all inferential knowledge is theoretical. If I see smoke, and I infer on that basis that there is fire, my inference isn't theoretical. It doesn't presuppose the truth of a theory. It assumes the truth of the statement "where there's smoke, there's fire," which merely describes an observable concomitance, and therefore doesn't qualify as a theory. But, supposing for argument's sake that there is unconscious mental activity and that we can have knowledge of it, the inferences that yield such knowledge *do* presuppose theoretical knowledge. If I infer from the fact that you are accident-prone that you have hostility towards yourself, I am obviously making several assumptions as to the nature of the psyche. Further, those assumptions don't merely register observable

concomitances, and they aren't independent of one another. Thus, those assumptions jointly constitute a theory. It follows that any knowledge that we have of the unconscious is of a theoretical character. In sections 7.0–7.1 we'll say more about the distinction between theoretical and inferential knowledge.

7.1 Is all knowledge theoretical?

Some epistemologists and cognitive scientists maintain that even perceptual knowledge is theoretical. This position may seem strange at first, but there is at least one reason to think that it may be true.

Even though *you* don't have to do any work to generate your own sense-perceptions, your mind and nervous system certainly have to do a great deal of work. Consider the disturbances of your sensory surfaces that lead to your sense-perceptions. If the surface of a rock or a piece of wood were similarly disturbed, no sense-perception would result. What is the difference between you and a rock? Your mind can, whereas a rock cannot, *process* the relevant disturbances in the right way. It's quite extraordinary that your mind is able to do this; for all it is given are various disturbances that, in and of themselves, bear no resemblance, or only an extremely abstract resemblance, to the things off of which the relevant photons (or whatnot) are bouncing. Nonetheless, on the basis of this meager information, your mind is able to generate perceptual representations that accurately depict those objects. The gulf between input and output—between, on the one hand, the surface-stimulations on the basis of which your mind generates its perceptions and, on the other hand, those perceptions themselves—bears a striking resemblance to the gulf between, on the one hand, the data on the basis of which a theoretical physicist posits the existence of a swarm of a certain kind of sub-atomic particle and, on the other hand, that particle-swarm itself. In the case of the theoretical physicist, that gulf can be bridged only by a theory. This suggests that the same is true of the gulf between disturbances of your eyes and the perceptions in which those disturbances eventuate.[189]

7.2 A priori versus a posteriori knowledge

Thus, one must have knowledge *in order* to acquire perceptual knowledge. Given that even a newborn can have sense-perceptions, it follows that

knowledge of some kind is coded into the innate structures of our minds.

Such knowledge is referred to as a priori knowledge. Thus, knowledge is a priori if it is coded into the very structures of our minds and therefore pre-exists any perceptions that we have.

If knowledge is *not* a priori, then it is referred to as a posteriori. Thus, a posteriori knowledge is *acquired* knowledge, and a priori knowledge that is within us, in some form, from the outset.

It is a matter of considerable debate whether there is a priori knowledge and, if so, precisely what it is that known a priori. (My own strongly held view is that there is a priori knowledge.) Among those who believe that we have a priori knowledge are Plato, Leibniz, Kant, and Chomsky. Among those who deny it are Aristotle, Locke, and Quine.

7.3 A second reason to think that perceptual knowledge is theoretical

Suppose that X is some stop sign. If you look at X from a certain distance—from, say, a hundred feet away—it will appear a certain way to you. Given only how X appears to you from that distance, you cannot tell whether you are looking at an object that is around 8-feet tall and a hundred feet away *or* an object that is 1,000-feet tall and much further away *or* an object that is 6 inches tall and very nearby. Nonetheless, you know, when you are looking at X, that it is not 1,000-feet tall and far away or 6-inches tall and very close, and that it *is* around 8-feet tall and about a hundred feet away. Given that X's appearance, by itself, does not reveal this to you, how do you know it? It is your background knowledge that, in conjunction with how X appears to you under the circumstances described, enables you to assign the right height to X. A continuation of our story may clarify this. Your friend Larry is right next to X. You can see the *comparative* heights of X and Larry; you can see that Larry is around three-fourths the height of X. So it is because of X's appearance *plus* this background knowledge that you are able to determine X's height.

This must be understood correctly. You can just *see* how tall, approximately, X is. But, for the reasons just given, it seems that background knowledge of the kind just described seems to be involved in, and to some

degree to constitute, your seeing this.

An argument similar to that just given in connection with seeing X's size can be given in connection with seeing any object's size. Indeed, similar arguments can be given in connection with seeing any fact about any object —in connection with seeing how fast an object is moving, what shape it has, how far it is from one, and so on. And extensions of those arguments suggest what is true of seeing facts about physical objects is true of *any* kind of perceptual awareness of *any* fact about them.

7.4. Inferential vs. theoretical knowledge

There is a reason to hold that not all knowledge is theoretical, namely that *many* inferences are involved in knowing a theory and, consequently, inferential knowledge is a precondition for theoretical knowledge. Some background points are needed to make it clear why, at least arguably, this holds.

First of all, not all inferences—not even all inferences that lead to correct conclusions—yield knowledge. If I infer that Bob has cancer on the grounds that he has red hair, my inference doesn't yield knowledge, even if, by coincidence, Bob does have cancer. The reason my inference doesn't yield knowledge is that the "inference rule" implicit in it—in other words, the principle that licenses it—isn't a good one. That inference rule is: "if a person has red hair, he or she has cancer." Since that rule obviously isn't "truth-conducive"—since, in other words, that rule isn't likely to yield correct conclusions, given correct premises—no inference made on the basis of it is legitimate. So for an inference that yields a correct conclusion to yield *knowledge,* one must make that inference on the basis of a legitimate inference rule.

Theories are *systems* of statements standing in various *dependence-relations* with respect to one another. (The word "statement" here is not meant to have any linguistic overtones and should be taken as shorthand for "truth or falsehood.") A dependence-relation holds between two statements if one of them supports the other—if, in other words, one of them provides some basis for inferring the other. Knowing that a dependence-relation holds between two statements involves either inferring the one from the other or,

failing that, recognizing that such an inference could be legitimately made. For example, my recognizing that there is a dependence-relation between "there is fire at time t in place p" and "there is smoke at t* in p*," where t* and p* are neighbors of t and p," involves *either* my knowing one of them on the basis of the other *or* it involves my seeing that such inference would be legitimate. In either case, I have inferential knowledge. (Even if I happened to see the fire and the smoke, and thus didn't have to infer "fire at t in p" from "smoke at t* in p*" or *vice versa,* my knowing that a dependence-relation holds between them involves my knowing that the one could be legitimately inferred from the other. And *that* knowledge, surely, deserves to be characterized as inferential.) Since, as we noted a moment ago, knowledge of dependence-relations precedes knowledge of theories, and since knowledge of such relations is inferential, it follows that inferential knowledge precedes theoretical knowledge.

Of course, those who deny that all knowledge is theoretical don't have this in mind. Usually those who hold that there is "direct" knowledge mean that there is no *inferential* knowledge—they use the terms "inferential" and "theoretically" interchangeably and therefore, since they're not interchangeable, loosely.

7.4.1 Inferential vs. theoretical knowledge (continued)

To be made cogent, the foregoing discussion must be qualified in a couple of ways. These qualifications will bring out some principles whose importance extends far beyond the points just made.

I said that a dependence-relation holds between two statements just in case one of them can be legitimately inferred from the other. Some might object to this on grounds like the following:

> Smith's crashing the car might have resulted from his being drunk. So the statements "Smith was drunk at t" and "Smith crashed the car at t*." But surely I cannot legitimately infer the second from the first. Given only that Smith is drunk, I can't infer that he'll crash his car. So, contrary to what you say, dependence-relations aren't always backed by legitimate inference rules.

First of all, it wasn't *merely* Smith's being drunk that led to his crash. It was his being drunk while being behind the wheel. (In fact, more than this is

involved. He wouldn't have crashed had the car in front of him not slowed down or if he hadn't been talking on his cell phone when this happened. It was a confluence of factors—his drunkenness, the behavior of the driver in front of him, his own distractedness—that led to the crash. Total causes are always confluences of circumstances. Thus, the dependence-relation in question holds, not between the two statements cited by the objector, but between "at t, Smith crashed" and "at immediately preceding time t*, Smith was driving while drunk and talking on his cell phone, when the car in front of him suddenly slowed down." The second statement *does* make it reasonable to infer that, at or around t, Smith crashed.

Admittedly, the second statement doesn't make it *certain* that he crashed. This is because the presumption that, given the various factors at play (his drunkenness and distractedness, the sudden deceleration of the car in front of him, etc.) the imminent crash could still be headed off. Maybe the person in the passenger-seat of Smith's car wrenched the steering wheel to the right, just in time to avoid the crash. Not likely, but possible. Maybe Smith is somebody whose driving is enhanced by drunkenness. Not likely, but possible. But even though, given the operative conditions, Smith's crashing cannot be known with certainty to occur, it doesn't follow that the inference from "Smith was drunk and chatting on his cell phone [etc.]" to "Smith crashed" isn't legitimate or, therefore, isn't licensed by a truth-conducive inference rule. What follows is that the inference rule in question is *defeasible.* An inference rule is defeasible if (i) it provides a reasonable *presumption* that, given certain premises, a certain conclusion holds, but (ii) that presumption may turn out to be false if other facts, not in evidence, happen to intercede.

There is a big difference between an inference rule that is defeasible, on the one hand, and one that is simply bad (i.e., not truth-conducive), on the other. A bad inference rule provides *no* reason for inferring one thing from another. A defeasible inference rule (when good) does provide such a reason. It doesn't provide as good a reason as a non-defeasible inference rule. But that doesn't mean that it's not good at all; goodness (truth-conduciveness) comes in degrees.

7.4.2 Rules vs. principles

There's also a big difference between a principle and a rule.[190] Rules are about outcomes. Principles are about tendencies. Rules have the form: "if this, then that." (If it's a raven, then it's black.) They're categorical. (All members of the category *raven* are members of the category *black*.) A rule that admits exceptions is false. (A single non-black raven falsifies the rule that all ravens are black.) A correct rule can be obtained by replacing a counter-exampled rule with one that is duly hedged with exception clauses. (So, supposing that Smith is the one non-black raven, then "all ravens other than Smith are black" is indeed a correct rule.)

But a duly hedged rule is not a principle. Principles are not watered down rules. It is no accident that so many drunk drivers crashed; it's not a statistical fluke. Nor is it parasitic on some other, more robust connection. In a society where the government forced blond people to live near nuclear waste, there would be a connection. Drunk driving is to crashing what being exposed to radiation (as opposed to being blonde in the society just described) is to getting cancer.

As we'll see in Chapter 17, empiricists think of principles as being watered-down rules—as being expressed by categorical statements to which there are exceptions. This isn't the right way to think about them. Principles are better thought of as registering the existence of defeasible forces. Being drunk can be thought as a car-crash-conducive force (or, more likely, confluence of forces) that, like most forces, can be overridden. Note that, whereas inference-rules are defeasible in a logical sense, forces are defeasible in a causal sense.[191]

8.0 Knowledge of ourselves and others

A person's knowledge of himself is obviously much better than his knowledge of others. You obviously have a much more accurate understanding of your own mind than you do of the other people's minds. Nonetheless, we can, within limits, have some knowledge of what others are thinking.

It is a matter of considerable dispute how we can acquire such knowledge. But here is a hypothesis that, I think, must surely be correct—at least for some, though perhaps not all, instances of knowledge of others.

We can, at least within limits, have direct knowledge of what is going on within our own minds. And we often know how we are likely to behave in response to an experience of a given kind. When others behave in a given way, I tend to ascribe to them the kind of mental state that accompanies instances of *my* behaving in that way. I know that when I am in pain, I tend to wince or clutch at the wounded part of my body (or some such). So when I see somebody else behaving in that way, I tend to assume that he or she is in pain. Comparable points hold with respect to our ascriptions of other kinds of mental states to others.[192]

8.1 Self-knowledge

People know much about themselves. I know that, right now, I have the intention of producing a helpful essay on the theory of knowledge. I know that, right now, I'm in a pretty good mood. I know that I enjoy playing the piano. I know that I wish that I could play as well as Arthur Rubenstein; I also know that I'm sad that I cannot play as well he could. All of you have comparable knowledge about yourselves.

A striking feature of at least *some* of our self-knowledge is how direct it is. My knowledge that I'm in a pretty good mood is even more *direct* than knowledge that I am now typing on a computer. But how is this possible? Isn't my knowledge that I'm now typing on a keyboard as direct as knowledge can be?

No. My knowledge that I'm typing on a keyboard is based on the sensory experiences I'm having. These experiences are both visual and tactile. ("Tactile" means "having to do with the sense of touch.") But, given only the nature of those experiences, it is a possibility, if only a theoretical one, that I'm dreaming or hallucinating.

My having accurate perceptions of, for example, the computer screen in front of me involves my body's being stimulated in very specific ways by causal processes that originate with that computer screen. (Light waves must bounce off of that computer screen and disturb my retinas.) Further, those light waves mustn't be distorted or altogether cut-off by intervening influences. If there were an uneven, thick piece of glass between me and the screen, those light waves would be distorted, and my perception of the screen

would be inaccurate. If there were a piece of thick wood between me and the glass, those light-waves would be cut off, and I wouldn't see the screen at all. So my having an accurate perception of the screen involves more than just my having a certain visual experience: that experience must be caused in a very specific way by very specific events in the external world (i.e., the world *outside* of my body).

A corollary is that an experience that is just like the one I'm having, but that is caused in the wrong way; it isn't a perception of the screen. Suppose that, using precision instruments, an extremely skillful surgeon is creating disturbances of my retinas that are exactly like the disturbances that *would* ordinarily be created by light bouncing off of a nearby computer screen. In that case, I will have an experience just like that of seeing a computer screen, even though I'm not seeing a computer screen.

There are typically many different ways of creating a given effect. If you want to hang a painting on the wall, there are a lot of different ways to do it. If you want to make it look as though Smith killed Jones, there are a lot of different ways to do *that*. One way, of course, would be to get Smith to kill Jones (by, for example, paying him to do so). Another would be to frame Smith. Thus, even if Smith is innocent, he can be made to *look* guilty. Similarly, even if there's no computer screen nearby, my retinas and, consequently, my nervous system, can be made to undergo the kinds of disturbances that would ordinarily be caused by light-rays that had bounced off of a computer screen.

Of course, remarks similar to those just made about visual experiences hold with respect to all other kinds of sensory experience. (I'll leave it to you to think of examples of this.)

Basically, sensory experiences, like forensic data, can be "forged"; they can be "fakes."

But such fakery seems impossible when it comes to at least certain kinds of self-knowledge. I can have an experience *just like* that of seeing a computer screen even if I'm not seeing a computer screen. But could I have an experience *just like* that of experiencing intense pain if I weren't experiencing intense pain? Surely not. After all, an experience just like that of intense pain *is* an intense pain. Of course, what we just said about pain is true

of many other psychological conditions.

Thus, where some psychological occurrences are concerned, the appearance and the reality merge. By contrast, where knowledge of the *external world* is concerned, the appearance and the reality are always distinct and are often (in fact, usually) separated by a large tract of space. (Even where touch is concerned, the object touched is remote from the events in your central nervous system that mediate your *experience* of the object.) Thus, from an evidential, and in fact a spatial, point of view, knowledge of at least some facts about one's psychological condition is considerably more direct than knowledge of facts about the external world—even when the externalities in question are literally being *seen* (or touched or heard, etc.).

8.2 Some limits to self-knowledge

Notwithstanding what was just said, there are reasons to believe that, in many cases, self-knowledge is as difficult to obtain as knowledge of the external world—that, in fact, knowledge of some facts about our own minds is as theoretical and indirect as our knowledge of sub-atomic events.

Before we can discuss these reasons, we must define a term and we must also make a distinction. The term to be defined is "consciousness." In everyday speech, that word is ambiguous. Sometimes, it simply means "knowledge" or "awareness." For example, in the sentence, "Smith's consciousness of the complexities involved in running a municipality is extremely limited," the word "consciousness" seems to have more or less the same meaning as the term "knowledge" or "awareness."

In philosophical contexts, the word "consciousness" has a different and narrower meaning. In such contexts, it refers to one's immediate experience. (Henceforth, we will use the term "consciousness" *only* in its narrow, philosophical sense.)

This must be understood aright. The term "consciousness" does *not* refer to the *objects* of your experience. Right now, let us suppose, you are feeling a cool breeze. The cool breeze is something that you are conscious *of.* In other words, it is an *object* of consciousness. But that cool breeze isn't identical with your consciousness or with any part of your consciousness. What *is* a part of your consciousness is the sensation that the cool breeze causes you to

have. What is also a part of your consciousness is your visual perception of the piece of paper that the cool breeze is blowing across your front lawn. But, to echo what we said a moment ago, the piece of paper itself is not identical with your consciousness or even with a part of it, even though it is something that you are conscious *of* (i.e., it is an *object* of your consciousness).

Here we must make a delicate point. Many constituents of our minds exist even when we are sleeping and even when, more generally, they are not occupying our consciousness in any way. You believe that 1 + 1 = 2. But it is only very rare occasions that this belief of yours is occupying your consciousness. Nonetheless, you obviously don't *lose* that belief when it's not occupying your consciousness. You still have that belief. This is made clear by the fact that it can readily make an appearance in your consciousness. For example, if I tell you that I'll give you an A in the class if you answer the question "what is 1 + 1?", you will without any difficulty be able to draw on your belief that 1 + 1 = 2, even though, in all likelihood, that belief was remote from your consciousness prior to my asking you that question.

So that belief, existing as it does when it is outside of consciousness, is not a *part* of consciousness. It is not, as we will henceforth put it, a *constituent* of your consciousness.

Thus, there are instances of knowledge that are not consciousness. And there are instances of consciousness that are not knowledge. The euphoria induced by a drug is an instance of consciousness; it is a component of somebody's immediate experience. But it isn't an instance of knowledge. (This is consistent with the fact that somebody experiencing such a euphoria could, and probably would, have knowledge *of* that experience, just as you have knowledge *of* the piece of paper blowing across your lawn).

8.2.1 An important fact about consciousness

One feature of consciousness (as philosophers use that term) is that, if something has consciousness, then there is "something it is like" to be that thing. If you are in pain, then there is "something it is like" to be you. So the presence of consciousness in something is sufficient for there to be "something it is like" to be that thing. Further, the presence of consciousness in something is *necessary* for there to be "something it is like" to be that

thing.

There is nothing it is like to be pinecone, since pinecones are not conscious. But if, by some miracle, a pinecone were to become conscious, then there *would* be something it was like to be that pinecone. The presence of consciousness in something is thus both *necessary* and *sufficient* for there to be something it is like to be that thing. In fact, consciousness may be *identified* with a thing's having the property there being something it is like to be that thing.

8.2.2 An important fact about your belief that 1 + 1 = 2

These points help clarify our earlier point that your belief that 1 + 1 = 2 is not a *constituent* of your consciousness (even though it may occupy your consciousness, just as the spectacle of litter blowing across a street may occupy your consciousness without, of course, being a veritable *ingredient* of that consciousness). Although there is something it is like to have a pain or a tickle, or to experience a sense-perception, there isn't anything it is like to believe that 1 + 1 = 2.

Of course, that belief may have *effects* on your consciousness—in other words, it may generate constituents of consciousness (without itself being such a constituent)—and there is something it is like to experience those constituents. For example, your having that belief may lead to your getting an A on a test, which in turn may fill you with joy.

8.2.3 A consequence of this important fact: the existence of the unconscious

Thus, your belief that 1 + 1 = 2 is not a *constituent* of your consciousness. It is not like a pain or a tickle or a perception. But it is clearly "conscious" in some sense of the word. (It is not buried in your subconscious; it isn't like the previously discussed processes by which your mind converts disturbances of your sensory surfaces into perceptions.) It is not a *part* of your consciousness. So in *that* sense it is *un*conscious. So to the extent that it is "conscious," it is in the sense that you are conscious *of* it.

What we just said about that particular belief is true of many, if not all, beliefs and, indeed, of many desires, aspirations, regrets, intentions, and emotions. Consider your intention to receive a degree from an institution of

higher learning. That intention obviously has effects on what you do. (You choose to show up to class, take exams, and so on.) And it therefore obviously has effects on what is occupying your consciousness. (Because of that intention, you are now reading a handout and are therefore having various perceptions and feelings that are constituents of your consciousness.) But that intention cannot be identified with any *constituent* of your consciousness. (That intention continues to exist even when you are asleep.) That intention is better thought of a structural fact about your mind that *underlies* and *regulates* the events that populate your consciousness, without itself being such an event. Similar remarks hold with respect to (for example) your affection for certain people, your dislike of other people, your confidence in your ability to (for example) play tennis, your doubts about your ability to (for example) play the harmonica, and so on.

These points give us insight into an issue of great importance. You don't *have* to be conscious of anything that you *are* in fact conscious of. You are now conscious of a piece of paper (or computer screen) in front of you. But you don't have to be, and in a few minutes you won't be. You may, at a certain moment, be conscious *of* the fact that you believe that 1 + 1 = 2. But you obviously don't have to be conscious of the fact. In fact, most of the time, you are not thinking about that fact and therefore aren't conscious of it (though you readily become so). So the truth is that we don't have to be, and in fact are usually not, conscious of much of what is in our own minds. This shows that, contrary to what a number of authors have alleged,[193] there is at least some reason to accept the thesis, which is urged by thinkers in the psychoanalytic tradition, that we are not conscious of much of what is in our own minds.

Of course, thinkers in that tradition hold not only that we *aren't* conscious of much of what occurs in our own minds, but also that emotional blocks may lead to our having very wrong and distorted and impoverished beliefs about our own minds. Whether that thesis is correct lies outside the scope of this book. But what seems reasonably clear is that there are good grounds for at least *one* contention of that school of thought—namely, that we are, at any given time, unaware of many of the events that constitute our very own mental lives. And, once this is granted, it becomes a garden-variety empirical question whether, unless we take special measures to learn about

ourselves, we have more than a superficial understanding of our own minds. [194]

9.0 Knowledge by acquaintance versus knowledge by description[195]

At the beginning of this chapter, we distinguished between two kinds of knowledge. The one consists of awareness of people (or, more generally, objects); the other consists of awareness of truths. There is reason to think that up to a point, but probably not entirely, the first kind of knowledge collapses into the second. The reason for this is that, with certain very specific exceptions, awareness of objects is identical with knowledge of truths of a certain kind—that, more specifically, it is identical with knowledge of truths to the effect that certain *descriptions* are in fact satisfied.

A story will help clarify what it means to say that a description is "satisfied." In order to collect insurance money, an unscrupulous store owner falsely claims that he is robbed. He gives a description of the (nonexistent) perpetrator of the (non-existent) crime. He goes to the police station and tells the police: "the person who robbed me was a tall male; he had a tattoo that said "I love mom" on his forehead; he had pointy gold teeth . . ." Here the store owner is really saying that there exists an object (a person) that has certain characteristics. To put it another way, he is saying that a certain description is satisfied, that description being "entity that is a human being who robbed me and has pointy gold teeth and is male and . . ." Any statement to the effect that a description is satisfied is known as an *existence-claim.* So the store owner is affirming the following existence-claim:

(1) There exists an object that is a *person, that robbed me, that is a male, that has pointy gold teeth, that has a tattoo that says "I love mom" on his forehead . . .*

That existence-claim is false, of course, the reason being that the relevant description (the underlined part of 2) is not satisfied.

But, of course, many existence-claims *are* true, the reason being that many descriptions *are* satisfied. For example, the following existence-claim is true:

(2) There exists a *person who is currently (in Feb. 2008) the U.S. president.*

And the reason that 2 is true is, of course, that the relevant description (the underlined part of 2) *is* satisfied. (It is satisfied by George W. Bush.)

9.1 Theoretical knowledge as descriptive knowledge

Whenever knowledge is theoretical or otherwise indirect, it is descriptive (i.e., it consists of knowledge of the truth of an existence-claim). We are aware of the existence of electrons. But we don't see them or otherwise sense-perceive them; and we certainly don't have the kind of direct experience of them that we have of some of our own psychological states (e.g., our joys and sorrows). So in what sense are we are of electrons? The answer: we know the truth of various existence-claims that they satisfy, for example:

(e) there exist various macroscopic, and therefore perceptible, phenomena that, given known physical laws, are most plausibly explained in terms of theories positing the existence of negatively charged particles orbiting around the nuclei of atoms . . .

We are aware of electrons in the sense that we know the truth of existence-claims that they satisfy. Since, as we discussed, an existence-claim is a statement to the effect that a description is satisfied, our awareness of electrons is entirely descriptive in nature.

Remarks similar to those just made about our knowledge of electrons hold with respect to our knowledge of anything theoretical or, indeed, anything whose existence is inferred, as opposed to known directly.

9.2 Perceptual knowledge as descriptive knowledge

Our knowledge of *directly perceived* objects is strictly descriptive. Consider your current visual perception. What it tells you is that a certain existence-claim is true, that existence-claim being along the following lines:

> (sp) There exists a piece of paper that is in such and such a location with respect to me, that has such and such a shape and color, that has such and such discolorations on it . . .

Perception is description. It is inconceivable that one should have a perception that didn't describe the world in a certain way and that did not, therefore, say of the world that it contained an object of having certain characteristics. (Even if you are looking into empty space, your perception is to the effect that the description *stretch of empty space* is satisfied by at least one entity in the world.) Thus, a given perception tells you that some existence-claim is satisfied, and perceptual awareness is therefore descriptive in nature.

This shows that awareness of sense-perceived objects, and therefore of other people, is identical with knowledge of truths of a certain kind (existence-claims), vindicating our earlier suggestion that knowledge of objects at least sometimes collapses into knowledge of truths.

Are there any cases where knowledge is *not* descriptive? The answer seems to be "yes." If you have an intense pain, it seems that your awareness of that pain is utterly direct and doesn't involve your knowing that some existence-claim is satisfied. The pain is right there, so to speak, in the theatre of your consciousness, and it needn't be represented to you through the intermediary of some description. So we have non-descriptive knowledge of at least some of our psychological states. But, for the most part, it seems that our knowledge is descriptive.

9.3 Procedural versus declarative knowledge

We have found that, at least arguably, there are *some* cases of awareness that are not awareness of truths. At least arguably, we can be *directly* aware of some of our own psychological states and can therefore be aware of them without being aware of truths that describe or otherwise implicate them. Are there other kinds of awareness or of knowledge that are *not* awarenesses of truth?

> Some philosophers and psychologists say "yes," the reasoning being as follows:
>
> (KH[196]) We must distinguish knowledge-that from knowledge-how. I know *that* Paris is the capital of France. This is knowledge of a truth. It is knowledge-*that.* I know *how* to ride a bike. This is knowledge, but it isn't knowledge of a truth. It is knowledge-*how,* not

knowledge-that.

> Any *skill* is a case of knowledge-how. It's easily shown that knowledge-how and knowledge-that are fundamentally different things. No matter how much I know about the physics and geometry of bike-riding, I may not be able to ride a bike at all; and no matter how good I am at riding a bike, I may not be able to tell you the first thing about the geometry or physics involved.
>
> No amount of knowledge-that adds up to a skill, and no amount of skill adds up to knowledge-that. So knowledge-how is not knowledge-that, but it is, of course, knowledge of some kind or other.

But it's a spurious argument for a false conclusion. Knowledge of a language is knowledge-how. To be able to speak English is to have a skill. But that skill involves having knowledge-that, for you must know various syntactic and semantic truths (e.g., the truth that "snow" picks out snow) to speak English. Knowledge of a language is operationalized semantic and syntactic knowledge. In general, procedural knowledge is operationalized declarative knowledge. (By "declarative knowledge," I mean knowledge of truths.)

Of course, not all procedural knowledge starts out as declarative knowledge. (Your ability to ride a bike didn't start out as non-procedural knowledge. It was procedural all along.) Moreover, one has no conscious access to the declarative knowledge internal to one's procedural knowledge. (One has conscious access only to analogues of that knowledge. A given piece of knowledge may be stored at many different cognitive levels. Knowledge is oftentimes "over-represented." A baseball player who is also a physicist may have two distinct pieces of knowledge of some one fact.) Contrary to what Ludwig Wittgenstein (1889–1951) and John McDowell (1942–) hold, the conclusion to be drawn from these facts is *not* that knowledge-that is no part of knowledge-how, and is instead that the mind is compartmentalized, and the right hand doesn't always know what the left is doing.

10.0 Intuitive versus discursive knowledge

Whenever we know something without knowing *how* we know it, we have

intuitive knowledge. Whenever we know something and we *do* know how we know, we have *discursive* knowledge. So to know something intuitively is to "just know it," whereas to know something discursively is to know it *while also* knowing the grounds on the basis of which one knows it.

Some short stories will clarify these points.

Story #1. Timmy, who is 10 years old, is in a classroom and he is looking at a figure on the blackboard. The teacher points to the figure and asks Timmy "what shape does that have?" Timmy correctly says "it's a circle." The teacher asks: "how do you know? Why is it a circle as opposed to oval?" Timmy says: "I just know. It's just obvious that it's a circle."

Story #2. Jones, who is a math major, is in a classroom and he is looking at a figure on the blackboard. The professor points to the figure and asks Jones "what shape does that have?" Jones says "it's a circle." The professor asks: "how do you know? Why is it a circle as opposed to oval?" Jones says: "because, if x and y are any two arbitrarily short segments of that shape's periphery, x's degree of curvature equals y's."

Commentary: Obviously both Jones and Timmy know that the figure on the board is a circle. But Jones knows something that Timmy does not know. But what, exactly? A possible answer is this: "Timmy doesn't know why the figure is a circle, whereas Jones does have that knowledge." But surely Timmy's belief is based on the very fact about the figure's shape that Jones has described. It seems to me that—in some way, at some level—Timmy is aware of the fact about that figure that Jones is describing. He doesn't know it in the same way as Jones—but he still knows it. So it is wrong, or at least very misleading, to say that Timmy doesn't know what it is about that figure that makes it a circle.

10.1 How intuitive and discursive knowledge differ

Here's what I think. Jones does, whereas Timmy does not, know *on what grounds he himself is judging the figure to be a circle.* Timmy and Jones both believe that the drawing is a circle, and (so we may assume) each believes it for the right reasons. But Jones knows, whereas Timmy does not know, what the considerations are that are moving him to make that judgment. The difference between Timmy and Jones is that Jones has a kind of self-

knowledge that Timmy lacks.

10.2 How intuitive and discursive knowledge differ (continued)

Here's another hypothesis (which, I think, is compatible with the one just given). Even though both Jones and Timmy know that the drawing is a circle, Jones knows this in *two distinct ways,* whereas Timmy knows it in only one way. Timmy has a strictly visual (or, in any case, sensory) knowledge of this fact, whereas Jones has what might be described as *propositional* knowledge of this same fact. Let us now develop this point.

First of all, a proposition is what is meant by a sentence—it is the information encoded in a sentence. The information provided by sentences has two striking features: first, it is very schematic; second, it is the kind of information that can be grasped only through an act of *judgment,* as opposed to perception. (These two facts are related.)

Suppose you see a certain man, whom you know to be Smith, next to a certain telephone pole, and you report that fact by saying:

(SP) "Smith is next to that telephone pole."

Compare what you actually *see* to SP. You don't just see *Smith.* You see an individual with very specific characteristics (a specific shape, a specific posture or series of specific postures, wearing specific clothes, having hair of a specific color, and so on). And you don't just see a *telephone pole.* You see an object having a specific shape, in a specific location, with a specific color (and with specific discolorations on it), with wires of a certain kind attached to certain parts of it, and so on. Thus, the sentence provides a very schematic and abstract representation of the information provided to you by your visual perception. So, from one viewpoint, the sentence *under*-represents the content of your visual perception.

But, from another viewpoint, the sentence *over*-represents the content of your perception. Given only what you see—given only what your eyes are telling you—it could be that you are seeing somebody other than Smith; it could be that you are seeing somebody who looks just like him but isn't him—or even that you aren't seeing a person at all and are in fact seeing a robot.

When you characterize that thing as "Smith," you are making a judgment as to who and what that thing is; you are making a judgment as to the biological species and as to the specific identity of that individual. Though *partially* based on what your eyes are now telling you, that judgment incorporates information that is not derived from that specific perception—information that is derived from past meetings with various individuals and from inferences drawn therefrom.

And given only what your eyes are telling you, you don't know that you are seeing a *telephone* pole. A telephone pole is something that serves a specific function within a complicated network of social and physical relations. When you are staring at the telephone pole, you cannot literally *see* it performing these functions. Thus, implicit in your characterization of that thing as a "telephone" pole is a judgment as to its social function. And, though partially derived from your sense-perception, the grounds for that judgment have roots in past experiences of yours and also in inferences drawn therefrom. So, from that viewpoint, SP over-represents what your eyes are telling you on this occasion.

Similar points hold with respect to the differences between Timmy and Jones. Jones isn't simply seeing a certain object's shape; he isn't simply uploading perceptual information. He is making a judgment about that information. (So Jones is generating "meta-information.") By contrast, Timmy is just uploading the information.

A related point is that, although both Timmy and Jones know that the drawing is a shape, Jones knows this fact in more ways than Timmy. This can be understood in terms of our earlier point that knowledge is descriptive. Here the fact known is that the drawing is a circle. Timmy knows one description of that fact—he knows a visual or graphic description of that fact. But Jones knows *both* this graphic description *and* a conceptual or propositional representation of it.

We thus seem to have two answers to the question "what does Jones know that Timmy doesn't?" First, Jones has a certain kind of self-knowledge that Timmy lacks: Jones does, whereas Timmy does not, know what the considerations are that led him to believe (rightly) that the drawing is a circle. Second, Jones knows the fact in question under two different descriptions (one pictorial, the other conceptual), whereas Timmy knows it only under one

description.

10.3 The paradox of analysis

These points may give us some insight into an old philosophical dispute. Consider the statement:

(CR) x is a circle if and only if x is a two-dimensional figure of uniform curvature.

Obviously CR is informative. It isn't like the statement:

(CC) x is a circle if, and only if x is a circle.

CR is an example of a *conceptual analysis.* A conceptual analysis is a statement that lays bare the structure of a concept.

According to G.E. Moore, it is paradoxical to suppose that conceptual analyses can be informative. There is, he says, a "paradox of analysis." Moore's reasoning seems to be this. The sentence "water is H_2O" expresses an analysis of water. In that sentence, the words flanking the "is" refer to the same substance. If they didn't, that sentence wouldn't express a correct analysis. For the same reason, the expressions flanking the "iff" in a sentence express a conceptual analysis must refer to the same concept. But if they did, then CR would be trivial, like CC, and thus fail to be an analysis.

This is not good reasoning. True—"water is H_2O" expresses a "chemical analysis"; and, in it, the two words flanking the "is" co-refer. True—CR expresses a "conceptual analysis." But it doesn't follow that, in it, the expressions flanking the "iff" must co-refer. Moore's reasoning is comparable to that of somebody who thinks that, since so and so isn't interested in having "legal" (law-related) discussions, so and so has no wish to have "legal" (law-permitted) discussions.

In any case, Moore's paradox is easily enough dissolved. The truth that one grasps in virtue of knowing that:

(i) x is a circle

is different from the truth that one grasps in virtue of knowing that:

(ii) x is a two-dimensional figure of uniform curvature.

Both of those truths describe the same *fact*—the fact that a certain object has a certain shape. But a given fact can obviously be described in different, but equally correct ways. The fact that x is a triangle can be described by saying that x is the area bounded by three straight edges, any two of which intersect, but not all of three of which do so. And that same fact can be described by saying that x is a three-sided, straight-edged, closed, planar figure. And:

(iii) x is a triangle iff x is the area bounded by three straight edges, any two of which intersect, but not all of three of which do so iff x is a three-sided, straight-edged, closed, planar figure

is just another way of saying that there are three different ways of describing x's shape.

11.0 Justification vs. discovery

The question:

(1) What events in Newton's mind led to his coming up with his theory of gravitation?

must be distinguished from the question:

(2) What merit does that theory have?

One can know the answer to the one question without knowing the answer to the other. To know why Newton had such and such beliefs, I don't need to know whether those beliefs are correct or otherwise meritorious, and to know whether those beliefs are correct or otherwise meritorious, I don't need to know why Newton had them.

Given a belief, the question "is it correct?" is to be distinguished from the question "how was it arrived at?" and knowing the answer to the one question doesn't necessarily involve knowing the answer to the other.

"Justification isn't discovery." Philosophers of science often say this. In so doing, they're affirming the points just made.[197]

11.1 The connection between justification and discovery

Granting that an idea's pedigree is to be distinguished from its merits, it's misleading to say that justification isn't discovery. To discover something is to come to know it. Knowing something involves having a justified belief in it. The justification for that belief will be encoded in the thoughts that led to that belief. No justification, no knowledge. No knowledge, no discovery. Thus, justification *is* discovery—or at least a good part of it.

We must distinguish *discovering* something from merely *coming to believe it.* One cannot discover something without coming to believe it, but one can come to believe something without discovering it. Coming to believe something doesn't necessarily involve being justified in believing it or, therefore, in knowing it or, therefore, in discovering it.

But to the extent that a person's thought-processes are expressions of intellectual ability, and not of prejudice or some other discovery-thwarting factor, that person's beliefs are likely to be justified and, consequently, that person won't come to believe something without being justified in doing so. Given that the conjectures that led to great scientific and intellectual advances are obviously reflections of intellectual ability, as opposed to blind luck, it follows that oftentimes the justification for such a conjecture *is* to be found in its psychological antecedents. It follows that justification is inherent, not just in scientific discovery, but also in scientific *proto*-discovery. In other words, it's inherent not only in the processes that lead to advances, but in the processes that lead to the conjectures that lead to such advances.

Great ideas often consist of many intricately interconnected subsidiary ideas. The odds of accidentally stumbling upon those ideas, as opposed to arriving at them on the basis of knowledge and insight, are exceedingly slim. Unless Newton and Keynes were just lucky, which is unlikely, the thoughts they had along the way to making their discoveries were driven by an understanding of how the evidence at their disposal ought to be interpreted. Those thoughts were therefore to the effect that, *since such and such evidence justifies thus and such contentions, thus and such contentions ought to be accepted; therefore, thus and such contentions are true.* So, setting aside

cases where the discoverer was simply lucky, the psychology and the logic of discovery mirror each other.

Pascal said "it is through intuition that we discover; it is through logic that we prove."[198] This is wrong. It is through logic that we discover and it is through logic that we prove. For there is no discovery without justification. An unjustified intuition isn't a discovery at all. And if an unjustified intuitions happens to be right, it's a matter of luck. While some non-trivial theoretical innovations may be made through luck, most of them are not. (I doubt that any of them are.) And a correct belief arrived at through intuition, as opposed to luck, is *ipso facto* justified. Proof is the identification of the logic embodied in discovery.

11.2 Discovery vs. prediscovery

"For all we know," many a philosopher of science says, "it was Newton's being hit on the head by an apple that led him to posit the inverse-square law. So we can't expect to find much of relevance to the justification of a conjecture in the processes that led to it. While it may be of interest to psychoanalysts, the psychological back story is no concern of those interested in the logical structure of science."

Not everybody who is hit on the head becomes an innovator in physics. So far as a blow to the head had anything to do with Newton's discoveries, it was only by way of its triggering an intricate train of thought. We must distinguish the instigators of a thought process from the thought process itself. Being hit on the head was no part of Newton's discovery process. It was, at most, a part of his discovery *pre*-process. The blow, supposing that it happened, merely activated the relevant processes and therefore wasn't constitutive of them.[199]

11.3 The roots of the problem

The connection between discovery and justification is easy to overlook if one doesn't bear in mind the following three points:

(1) Having a justification for a belief doesn't always involved knowing what that justification is.

(2) Knowing that justification doesn't always involve being able to say what it is. (One can't always articulate what one knows.)

(3) Being able to articulate that justification is different from actually doing so and, therefore, from transmitting that justification to others.

11.3.1 (1) explained

There is often a delay between having a justification for a belief and knowing what that justification is. My intuition tells me that so and so is ill-intentioned, even though, on the surface, he is behaving amicably; that a given sentence is grammatical, even though my grammar book classifies it as ungrammatical; that a given animal is about to attack, even though it isn't giving off any of the obvious signs.

We trust our intuitions, and our confidence in them is usually vindicated. (Speaking personally, I always regret second-guessing my intuitions.) But when we have them, we don't know *why* they are to be trusted. If we did, they wouldn't be intuitions, since an intuition is, by definition, a belief whose justification, if it exists, is unknown to the person having the belief.

Some have said that, for this very reason, intuitions are never knowledge. "If one has grounds for believing something," it is said, "one knows what those grounds are."

This reasoning couldn't possibly be more off the mark. It is only to the extent that one hasn't fully internalized the relevant principles that one's thought *fails* to be intuitive. It is only pedants and neophytes who rely on discursive thought. Good pianists don't have to *think* before they hit every note. Bad pianists do. Good pianists, unlike their bad counterparts, have internalized the relevant rules and, for that reason, don't need crutch of conscious, discursive mentation. Relying no consciously executed operations to apply rules is like relying on an instruction manual.

Discursive thought is remedial thought, and thought is non-remedial only insofar as it's intuitive. It doesn't follow that all non-discursive is good. But it does follow that not all intuitive thought is bad. Given that knowledge is intuitive only in so far as its possessor doesn't know the justification for it, it follows that one can be justified in having a certain belief without knowing what that justification is.

Nonetheless, philosophers of science tend to hold that one has a

justification only when one states, or at least can readily state, one's justification.[200] For this reason, they're forced to say intuitions are never justified. Given that scientific breakthroughs result largely from intuitive thought and often themselves take the form of intuitions, philosophers of science have no choice but to see the generation of discoveries as having no connection, or only a fortuitous connection, to the eventual justification of them.

11.3.2 (2) explained

One can know one's reason for having a given belief without being able to state them. A skillful interrogator may know why he believes that the suspect his lying but not able to say why he thinks this. His judgments are made on the basis of subtle cues that he might have a hard time articulating. Articulating one's knowledge may involve translating it from one medium into another. This is the case with the interrogator. His judgments are responses to visual information. Articulating those judgments would involve translating that information into a non-visual medium.

Imagine the following. Max believes that metal usually expands when heated, and this belief of his is thoroughly justified. (He's done numerous experiments, absorbed copious amounts of theory, and so on.) But for some reason Max refuses to *state* his reasons for having that belief. Max's situation is very different from the interrogator's. Unlike the interrogator, Max knows what the justification for his belief is. In any case, even though Max's belief is thoroughly justified, it could be said that Max has "failed to justify" it. Thus, the expression "unjustified belief" is ambiguous. On the one hand, it can refer to beliefs that aren't justified. On the other, it can refer to beliefs the justifications of for which, if they exist, haven't been stated.

11.3.3 (3) explained

Even though a given scientist has cogent reasons for accepting a given hypothesis, and makes every effort to communicate them to others, it doesn't follow that others know what those reasons are. Others may not understand what that scientist is saying—maybe his theories are too advanced for them. Maybe they do understand what he's saying but haven't yet a chance to authenticate his results. Until they have good reason to accept the scientist's testimony, *his* justification for accepting the theory in question isn't yet

theirs.

When epistemologists say that a belief is justified, they mean that *the person who has it* is justified in having it. When philosophers of science say that a belief is justified, they mean *either* what epistemologists mean *or,* what is more likely, they mean that people *in general* are justified in having it.

There was a time when many people believed that smoking causes lung cancer, but when only a few researchers actually knew it, since those researchers alone knew the relevant clinical data. During that time, it would have been correct to describe "*the* widespread belief that smoking causes cancer" as unjustified.

In general, there are many cases where *the* belief in such and such is unjustified and therefore isn't knowledge, even though specific scientists had already discovered such and such and, therefore, knew it and, therefore, were justified in believing it. It thus happens that such and such is discovered before *the* belief in it is justified. It's obvious why this might lead one to hold that discovering might precede justification and thus have nothing to do with it, and this is exactly what philosophers of science were led to hold. But it's the wrong conclusion to draw. No one person can discover something without knowing it or, therefore, be justified in believing it. The scientists who discovered that smoking causes cancer were justified in believing it. Their non-scientific contemporaries were *not* justified in believing it—but they didn't know it either. And when in due course they did know it, it was only because they were justified in believing it. Nowhere in all this do justification and discovery pull apart.

Chapter 11

Cartesian Skepticism and the Birth of Epistemology

1.0 Descartes' importance to philosophy

Pre-modern philosophy tended to focus on two questions: (i) Does God exist? (ii) Do universals exist independently of thought? A universal is a property or characteristic.

Modern philosophy is generally held to have begun with the work of Rene Descartes (1596–1650). Descartes believed in God's existence, and he also believed that he knew exactly how God's existence was to be proved. He therefore considered the case to be closed on question (i). Descartes wasn't interested in question (ii) and said little about it.

Descartes' single-handedly invented the discipline of analytic epistemology. But Descartes' specific epistemological views aren't what make him so important to philosophy. What did so was his conviction that we shouldn't believe anything that cannot be rigorously established. For Descartes, *nothing* ought to be taken on faith. Everything, including God's existence and the legitimacy of scripture, had to be established. Even though he himself believed in God, he felt it necessary to *justify* his belief in God. And did he justify it, or at least tried to, by providing two carefully constructed argument's for God's existence, which we'll consider in a few pages.

Thus, the central tenet of Descartes' work is that nothing ought to be accepted unless there is a logical, as opposed to faith-based, reason for doing so. So far as this was even *a* tenet or pre-Cartesian philosophy, it wasn't the main tenet of it. But it *is* the main tenet of modern philosophy, which is why Descartes is generally seen as the founder of it.

In addition to being a great philosopher, Descartes was a great mathematician. He invented the discipline of analytic geometry. This fact about him is deeply relevant to his philosophical work. Descartes believed that *all* knowledge ought to be modeled on geometrical knowledge—more specifically, on Euclid's ***axiomatization*** of geometry.

Euclid attempted, with some success[201], to show that every accepted principle of geometry could be derived from a small number of assumptions that he referred to as "axioms" or "postulates." (The distinction between axioms and postulates isn't important, and these days only the term "axiom" is used.) As contemporary logicians and mathematicians use it, the word "axiom" refers to any statement that, in a given argumentative context, is simply assumed to be true and therefore isn't argued for. But Euclid used this word (or its Ancient Greek counterpart, rather) to refer to what he believed to be self-evident truths—truths that could not be grasped without being seen to be true and that neither required nor admitted of justification. (Among Euclid's axioms are "all right angles are equal to each other," "any two points are connected by some straight line," and "a given line segment can be extended *ad infinitum*.")

Descartes tried to show that what Euclid did for geometrical knowledge could be done for *all* knowledge. That is, he tried to show that human knowledge could be derived from a few self-evident truths. This was one of the objectives of his most famous work, the *Meditations on First Philosophy* (or simply "the Meditations").

The other objective of that work is to refute *skepticism.* Skepticism is the doctrine that we can't establish that any of our sense-perceptions represent the world as it really is—that we can't know that our experiences aren't just one big hallucination or dream.

I believe that I am now typing on a keyboard. But how do I know this? This question might seem trivial. But it turns out not to be so easy to answer. Descartes tried to answer it. He failed. He also failed to axiomatize human knowledge. But he made some good points along the way. Plus, many a correct and important epistemological view is to be understood *in terms of* Descartes' views, which is why, even though the latter are often wrong, they deserve careful consideration.

Even though the *Meditations* are concerned primarily with the justification of our beliefs, they also concern the existence and nature of God, the roles played by reason and sense-perception in the acquisition of knowledge, and the relation of mind to body. Descartes felt it necessary, with some justice, to deal with these topics in order to deal with the epistemological problem with which the work is primarily concerned. Below are summaries and critical evaluations of the meditations.

2.0 Meditation 1: "Of the things of which we may doubt"

How do I know that I have two hands? The obvious answer is: I can just *see* that I do. But how do I know I am not hallucinating or dreaming? This question is not so easy to answer.

Historically, there have been two ways of answering this question: the *foundationalist* way and the *coherentist* way. Aristotle was a foundationalist, and so was Descartes.[202] According to foundationalism, all knowledge coincides with, or rests on top of, knowledge of certain self-evident principles. So the structure of a person's knowledge is like that of a building. There is a ground floor, or foundation, on which everything else rests. Different foundationalists have different views as to which principles are foundational. We'll discuss Descartes' views on this in a moment.

Coherentists say that there are no foundational pieces of knowledge; there is no ground floor, so to speak— no self-evidently correct beliefs with which either all knowledge coincides or on which it rests.[203] Coherentists say that knowledge is what results when, given various *data* that are *not* knowledge—for example, various sensations or perceptions—*if* those data support one another in certain ways, then certain reasonable inferences can be made on their basis, those inferences leading, in favorable cases, to knowledge. For example, right now, I am having various sense-perceptions (or, in any case, various experiences that I take to be sense-perceptions). I see (or think I see) a table and computer in front of me, etc. I also have various other perceptions (e.g., I feel a certain resistance every time I push one of the buttons on the keyboard that, so I believe, I am using to write this document).

I have had various similar experiences in the past. For example, visual experiences of a certain kind (e.g., experiences as of seeing keyboards) have been accompanied by tactile experiences of a certain kind (e.g., experiences as of touching keyboards); and I've noticed that experiences of the one kind are connected, in regular and predictable ways, with the experiences of the other kind. This point *mutatis mutandis* holds of our sensory experiences in general. So my various experiences, taken jointly, form a coherent and consistent picture of the world; and for that reason, I can reasonably infer that there is an external world of which those experiences are a more or less faithful picture, albeit a very incomplete one.

For the coherentist, no perception by *itself* constitutes knowledge. Only vast collections of experiences, supplemented by inferences, do so. For the foundationalist, on the other hand, there are some self-evidently correct beliefs—some beliefs that cannot be had without *ipso facto* having knowledge.

Descartes is, ultimately, a foundationalist, although towards the end of the *Meditations* he makes some coherentist points. Aristotle is a hard-core foundationalist. (See the first five pages of the *Posterior Analytics.)* Since around 1950, coherentism has been more popular than foundationalism, even though there are some contemporary foundationalists (e.g., Roderick Chisholm, who died in 1998, advocated a very compelling form of foundationalism—one which, incidentally, is not unlike Descartes').

In any case, when trying to answer the question "how do I know that my perceptions are not hallucinations?" Descartes does not, at least not in *Meditation 1,* consider the possibility that a coherentist answer might be correct. He says that what he must do to answer that question is to find "indubitable" principles on which all knowledge rests. This brings us to:

3.0 Meditation 2: "Of the nature of the human mind; and that it is more easily known than the body"

Descartes says that certain beliefs about our own minds are self-evidently correct, but that the same can never be said of our beliefs about external

objects. First of all, Descartes says, whereas I can doubt that there is an external world, the reason being that I might be perpetually hallucinating or dreaming, I cannot doubt that my mind exists. After all, I can doubt the existence of my own mind only if I have a mind: there is doubt only where there are minds, since doubting is a form of mental activity. (Author's commentary: This point seems spot on.) So any doubt as to the existence of mental processes is self-undermining. But doubt as to the existence of the external world is not self-undermining: given only that I *seem* to see a stick, it doesn't follow that there really is a stick there (for, as we said before, I could be hallucinating or dreaming).

In another work (not in the *Meditations*), Descartes famously says "I think; therefore I am." We have just seen the meaning of this. One cannot think without existing. Since doubting is a form of thinking, and since thinking involves existing, one cannot doubt one's existence without existing. Therefore, doubt as to one's own existence is incoherent. This gives Descartes an Archimedean point on which to build the structure of human knowledge: he knows that he exists; and, so he believes, he can on that basis justify all of the beliefs that we pretheoretically regard as true (e.g., that there is an external world with people and rocks and dogs in it).

Descartes makes another good point about the difference between knowledge of self and knowledge of the world. As just noted, given only that you have an experience *like* that of seeing a stick, it doesn't follow that there really is a stick there. But given only that you have an experience *like* that of being in pain, it *does* follow that you really are in pain. After all, an experience *like* pain *is* pain, whereas an experience *like* that of seeing a stick isn't necessarily an experience of seeing a stick. Descartes seems to hold that what we just said about knowledge of pain is true of all cases of self-knowledge. This is, in my view, a gross over-statement. But we don't need to pursue this, since it isn't central to Descartes' main project; and Descartes' core point (that at least certain kinds of self-awareness couldn't be delusional) seems correct.

Descartes makes another, extremely clever point. He says that, although we *think* that our basic beliefs about the physical world are purely perceptual, those beliefs are actually replete with inferential, even theoretical, content. Imagine the following. There is a piece of wax in front of you. It is solid and

has a definite shape. You put it on a warm pan. It slowly loses its shape. You are watching it the whole time. After 10 minutes, the wax doesn't really have a shape, and is just an amorphous puddle. Nonetheless, you know, after the 10 minutes have passed, that the puddle you are looking at is identical with the hexagonal solid object that you were looking at 10 minutes earlier. This knowledge, Descartes says, cannot be *purely* perception-based. For, if you stick strictly to what your senses are telling you, you would hold that the waxy puddle now in front of you is not identical with the previously seen solid object, your reason being that the puddle and the hexagon have different shapes. Therefore, your identification of the puddle with the solid object must involve a certain theoretical or inferential component: it involves your having the belief that an object does not lose its identity solely on account of changes in its shape—that is, that certain kinds of transformations undergone by objects (e.g., changes in temperature, shape, or fragrance) are consistent with the continued existence of those objects. *That* piece of knowledge is not perceptual and is better characterized as *meta*perceptual or as *pre*perceptual, since it is a *pre*condition for your acquiring knowledge from your senses. For if you didn't have it, then every time you saw an object change, you would falsely believe that it had been obliterated and then replaced by an other, similar object. But that isn't how it works: you believe that you are seeing some *one* object undergo change; and that is why you form the correct belief that the world consists of objects that, although constantly changing, do so more or less continuously, as opposed to the false belief that, at every moment, everything is obliterated and immediately thereafter replaced with new objects. (Author's commentary: I am touching up Descartes' argument a bit, so as to make it continuous with our discussion of rationalism. But the touch-up is more or less consistent with Descartes' original argument.)

Thus, any belief about the external world involves an element of inference, an element of theoretical extrapolation, and is to that extent precarious: after all, the more inferences you make, the more likely you are to be wrong. By contrast, one's knowledge that one is sad doesn't involve any inference and is thus *not* precarious. So our beliefs concerning our own minds —concerning their existence and natures—are considerably more solid and reliable than those concerning the external world. And those beliefs constitute the core of self-evident truths on which, in Descartes' view, our knowledge

of everything else (apart from purely conceptual truths; e.g., *1 + 1 = 2*) rests. To clarify the parenthetical remark: Descartes is a hard-core rationalist and thus believes that not all knowledge is derived from the senses. Indeed, as we just saw, he thinks that some knowledge is a precondition for perceptual knowledge. He also thinks that it can be known, without having to any empirical work, that triangles have three sides, that causes precede their effects, etc. Descartes' foundationalism
is strikingly modern and resembles Chisholm's quite a lot. (I happen to accept Chisholm's foundationalism, at least with some reservations.)

4.0 Meditation 3: "Of God, that He exists"

Although Descartes believes that some beliefs about oneself are self-evidently true, he doesn't think that our common-sense beliefs about the external world, which he believes to be true and which he wishes to justify, can be justified solely on the basis of our self-knowledge. Descartes thinks that, in order to justify our beliefs about the external world, it must be supposed that God exists and that, moreover, God is all-good (and would, for that reason, guarantee that our beliefs are more or less on the right track, at least most of the time).

In this Meditation, Descartes provides an argument for God's existence. (In Meditation 5, he provides a different argument.) The argument given in Meditation 3 is much better than the one given in Meditation 5.) Here it is.

The cause of an event must have "at least as much reality as the effect." The idea seems to be that something cannot come from nothing. If you make a statue, you're not going to create new clay (or marble); you're simply going to redistribute the clay that was there to begin with. (Everything comes from pre-existing materials: the quantity of matter in the universe is preserved, and all changes involve redistribution in that matter, as opposed to creation *ex nihilo* of new matter. "*Ex nihilo*" means "out of nothing.") Similarly, events consist in redistributions of objects, not in creation *ex nihilo.* (Author's commentay: This point has been denied, and perhaps it is false; but it is obviously reasonable enough.)

Descartes actually goes on to affirm a very strong version of the principle that something never comes from nothing. He says that effects must not only consist of the same material as their causes but must actually *resemble* their

causes. The idea seems to be that the hotness of the pan resembles the hotness of the flame which made the pan hot and, further, that *all* cause-effect relations can be understood on this model. (Author's commentary: This doesn't seem right. Redistributions in mass-energy often lead to states of affairs that have little or no resemblance to their predecessors. Mozart's 40th symphony resulted from various mental and neural events, but it doesn't resemble those events very much, except perhaps in some extremely abstract sense. Effects are often wildly different from their causes.)

Here is the next step in Descartes' argument. Each of us has concepts of various forms of infinity—of infinite goodness, infinite intelligence, etc. That means, says Descartes, that infinite goodness, intelligence, etc., must *in some sense* be in each of us. (Author's commentary: This is where I think Descartes goes totally wrong. I can have an idea of a trillion dollars without a trillion dollars being "in" me, in any significance sense. Similarly, I can have an idea of infinite intelligence without being infinitely intelligent myself.)

Descartes then combines the points just made. He says that since

(i) something never comes from nothing,

(ii) that effects resemble their causes, and

(iii) that there is, in some way or other, infinite goodness, intelligence, etc., in each of us,

it follows that

(iv) there really must be an infinitely intelligent, good, etc., being in the world—that, in other words, there must really be a God.

5.0 Meditation 4: "Of the true and the false"

Having, so he believes, established that a good God exists, Descartes has, so he believes, made it possible to establish that most of our beliefs are correct. The argument here is straightforward. Since God is good, he wouldn't want us to be wrong all the time. So God has given us the power to generate correct beliefs and he has set up the world in such a way that our belief-forming mechanisms (our senses and our methods of inference) tend to generate a correct picture of the world.

But here, as Descartes points out, we run into a problem. Even though we

seem to have some correct beliefs, we're also disastrously wrong a great deal of the time. If God is so well-disposed towards us, why would he let us have so many wrong beliefs?

Here Descartes provides a rather Byzantine argument, which may be summarized as follows. God has given us free will. That was good of him to do, of course, since we'd rather be free than be robots. Next step: We arrive at wrong beliefs only when we abuse our free will. Just as a person may abuse his/her free will by doing drugs, so he/she may abuse it by drawing inferences without sufficient evidence. So, being good, God had no choice but to give us free will. But since we have free will, some of us will inevitably abuse it and make wrong inferences (and also do other bad things).

But, says Descartes, there is a bright side to this story. If we don't draw conclusions that are not warranted by the evidence, we'll have only correct beliefs. Here Descartes ends up putting forth an argument to the effect that most of our beliefs about the external world are correct, and that argument has both foundationalist and coherentist elements. (It's a pretty good argument, I think, even though two of the premises are clearly false.) First of all, we have self-evidently correct beliefs about our mental states (as we discussed a while ago). Second, something never comes from nothing and, moreover, effects resemble their causes. So the perceptions that we have must have causes and, moreover, they must resemble those causes. Further, Descartes says, those causes don't lie within us. How do we know this? Because, Descartes says, one's mind is completely transparent to its owner; and when a given person looks inside himself, he doesn't encounter anything that *resembles* the perception he is having of (say) the table in front of him—he just encounters the perception. So the causes of our perceptions must lie in the external world. Thus, our perceptions are accurate, at least most of the time, the same being true of the beliefs based thereupon.

Why do some unfortunates have hallucinations, etc., despite God's infinite goodness and his desire for our perceptions to be correct? As Descartes realizes, this question cannot be answered by saying "because those people have abused their free wills," since hallucinations are involuntary. Descartes' way of dealing with this problem is *ad hoc* and unconvincing; so we'll skip it.

5.1 Evaluating Descartes' argument

Descartes' first premise (*minds are completely transparent to themselves*) is unqualifiedly false. (See Chapter 10 for a discussion of why this is so.)

The second premise (*effects always resemble their causes*) is also false. How does your pushing the alarm button resemble the ensuing noise?

But it might be possible to fix up this premise. My pushing the button is a highly indirect cause of the bell's ringing. The shorter the series of events linking (distal) cause and effect, the more likely cause and effect are to resemble each other. It may even be (though this is a very strong claim) that, for any degree of resemblance R, short of 100%, there is some length L such that, if the series is cut down to L or less, output will resemble input to degree R. So even though Descartes is not right to say, without qualification, that effects resemble their causes, a closely related point may be correct, namely: effects *tend* to resemble their causes.

Twentieth- and twenty-first-century philosophers ridicule Descartes for assuming that effects resemble their causes. I myself think it is in this point that an adequate refutation of skepticism lies. This cryptic assertion is clarified and defended in Chapter 12.

6.0 Meditation 5: "Of the essence of material things; and, again, of God: that he exists"

The main point of this Meditation is that there is a second proof of God's existence. This second argument coincides, nearly enough, with Anselm's "ontological argument." Here it is. God is maximally excellent, by definition. Anything bad or ugly or stupid would, for that reason, not be God. In general, anything that is deficient in any way cannot be God. Not existing is a way of being deficient. So God exists.

In Chapter 25, it is explained why this argument fails.

7.0 Meditation 6: "Of the existence of material things, and of the distinction between the mind and body of man"

Here Descartes puts forth his (in)famous argument for *mind-body dualism,* the doctrine that mental entities (e.g., beliefs) are not identical with physical entities (e.g., patterns of neural stimulation). Here is that argument.

Let x be a belief. Suppose that I rightly believe x to be a belief. In other words, suppose that x has the property of being believed by me to be a belief. For any pattern of neural stimulation y, it doesn't follow that I believe that y is a belief. That is, it doesn't follow that y has the property of being believed by me to be a belief. So x and y don't have the same properties and therefore cannot be identical.

The customary response to this argument is to say that it involves an "intensional fallacy." Benjamin Franklin was the first postmaster general and he was also the inventor of bifocals. So the first postmaster general = the inventor of bifocals. But one can believe that the first postmaster general = the first postmaster general without believing that the first postmaster general = the inventor of bifocals. And from the fact that one can have the first belief without having the second, it obviously doesn't follow, and it isn't the case, that the first postmaster general ≠ the inventor of bifocals. For the same reason, from the fact that believing that belief x = belief x doesn't entail believing that belief x = pattern of neural stimulation y, it doesn't follow that x ≠ y.

The idea is that, in certain sentential contexts, Leibniz's Law breaks down (or, rather, appears to breakdown), and that verbs such as "believe," "know," and "doubt" create such contexts. Such contexts are referred to as "intensional" contexts. A context that is not "intensional" is "extensional." So a context is "intensional" if it is one where there can be an apparent violation of Leibniz's Law, and a context is otherwise "extensional." Why, exactly, intensional contexts are created by certain expressions is an extremely delicate question, as is the related question of how these apparent breakdowns are to be reconciled with the fact that there are surely no actual exceptions to Leibniz's Law, since Leibniz's Law is identical, ultimately, with the obviously correct principle that nothing can fail to have a property that it has. But in this context we may set these subtleties aside. (They are thoroughly discussed in Chapter 6.)

Given only that Descartes' argument fails, it doesn't follow that mind is body. But it does mean that Descartes' argument must either be augmented, or replaced with a different argument, *if* it is to be established that mind and

body are distinct. Efforts in this direction have been made in recent years, some by yours truly (although I now believe that beliefs *are* patterns of neural stimulation). The relationship between mind and body is the topic of the rest of the present chapter.

8.0 Different views as to the relation between mind and body

Beliefs, hopes, tickles, pains, aspirations, dreads, intentions, fantasies, and longings are *mental* entities. Chairs, explosions, rocks, gusts of air, and quagmires are *physical* entities.

How is the mental related to the physical? Here are the different views:

(1) Dualism: mental entities are *not* physical entities.[204]

(2) Materialism: mental entities *are* physical entities.[205]

(3) Dual aspect theory: the mental and the physical are different aspects of the same thing.[206]

(4) Mental eliminativism: there is no such thing as the mental. There are brain-states—but no beliefs, tickles, etc.[207]

(5) Idealism: there is no such thing as the physical. There are experiences, hopes, etc. but no chairs, rocks, etc.,[208]

Right now, let us focus on (1) and (2), since they are, pretty obviously, the only two of these five views that have any real chance of being correct. Later we'll discuss why, despite their obvious falseness, people have held (3)–(5).

First of all, (2) is *not* the view that mental entities don't exist. (2) must be distinguished at all costs from (4). According to (2), perceptions, beliefs, doubts, and so on, all exist—but they are identical with brain-states. The materialist says that, just as water is H_2O, so your belief that 1 + 1 = 2 is identical with some pattern of neural stimulation (or some such). Just as the physical chemist isn't *denying* the existence of water in identifying it with H_2O, so the materialist isn't *denying* the existence of the mental in

identifying it with physical entities. Although there may be *some* tendency on the part of materialists to be dismissive of certain features of the psychological realm, they don't *have* to be thus dismissive: Chomsky is a materialist, as was Freud—and those two were nothing if not great believers in the psyche.

Advocates of (2) typically (but not universally) hold that mental entities are identical with brain-events and brain-structures. For obvious reasons, they don't usually hold that mental entities are identical with trees or rocks or anything else that is not a part of one's body (or, more specifically, one's central nervous system).

I have published some papers arguing for dualism. I am now a materialist. But I am still fundamentally open on this issue.

Here are my own views. I believe (as I just said) that materialism is correct. Second, I believe (very strongly) that *functionalism* (a very popular form of materialism: to be defined later) is wrong. Third, although I believe that materialism is correct, I don't think (paradoxical though it may sound initially) that it can ever be explained *why* the mental is identical with (some aspects of) the physical.

Right now, let us focus on some of the basic arguments for and against (1). (In the process of discussing these arguments, we will find that there are different versions of both materialism and dualism.) Of course, since (1) is the negation of (2), any argument for (1) is an argument against (2), and any argument for (2) is an argument against (1).

9.0 Arguments for dualism

First argument

The mental obviously has properties not had by the physical. For example: mental entities are representational; they are purposive; they have "phenomenology" (i.e., there is something it is like to have them). Since physical events obviously don't have those properties, physical events aren't mental. (This argument was given by Franz Brentano in *Psychology from an Empirical Standpoint.*)

This argument evaluated

Though it has force, there is a problem with this argument. The materialist can say that physical events *do* have these properties, but that this simply isn't obvious to introspection. For example, physical events have lots of properties that aren't obvious. When you look at a chair or a rock or a pool of water, your perception doesn't tell you anything about the microstructural properties of what you are seeing—but it would obviously be absurd to conclude that what you were seeing *didn't* have invisible (to the naked eye) microstructural properties. Similarly, says the materialist, when you introspect a pain or a tickle or a belief, your introspection doesn't tell you anything about the neural properties of what you are introspecting—but it would be absurd to conclude, on that basis alone, that what you were introspecting *didn't* have neural properties.

A counter-argument

Some mental entities seem to be *appearances.* Consider your perception of (e.g.) some nearby table. Your perception is just the event of something's *appearing* to you. It doesn't make sense to say that there is more to an appearance than meets the eye. After all, what doesn't meet the eye is *ipso facto* not an appearance. Your perception of the table obviously doesn't *seem* to consist of lots of neural events. So if it *did* consist of such events, they are beneath the surface—they don't "meet the eye" of introspection. But given what we said a moment ago, this would mean that such events *aren't* part of that perception.

This counterargument evaluated

I'm genuinely open as to whether this argument is a good one. I believe (for reasons that I will follow) that materialism is correct. But I cannot find any errors in the just-stated counter-response (though others claim to be able to do so). If you can identify some consideration that resolves the matter, or that gives some weight to a certain (attempted) way of resolving the matter, I would be extremely interested in what you have to say.

Second argument for dualism—the explanatory gap

Consider some psychological statement; for example,

(P) Last year I was very sad, because I completely failed in my efforts to write a best-selling novel.

(P) is obviously the kind of statement that could be true. But can the truth of (P) be understood in terms of facts about neural-firings? Can we conceive of some statement about a person's *body* that made it clear why (P) was true? Of course, we can imagine discovering that people feel sad when, and only when, certain neurons in their brains are firing. But would such a discovery really give us any insight into the nature of the dependence (if any there be) of that psychological fact on the existence of those neural states of affairs?

Let us generalize this point. There is no denying that we can discover psycho-physical *concomitances;* that is, that we can discover truths of the form *mental phenomenon m occurs when, and only when, physical phenomenon p occurs.* But there always seems to be a "gap" between the neural event and the associated psychological event. Permit me to clarify this last point.

Water is H_2O. Of course, when you look at a glass of water, you cannot *see* that it consists of H_2O molecules. But suppose that you could shrink yourself down to the size of such a molecule. In that case, you could literally *see* the molecules with which physical chemists identify water.

When you look at a brain, you see beige tissue—you don't see emotions, beliefs, etc. If you were to shrink yourself down the size of a molecule, you *still* wouldn't see emotions, beliefs, etc. In other words, you *still* wouldn't see

the entities that the materialist identifies with brain-states.[209] You would see the entities with which the *physical chemist* identifies brain-states (atoms, molecules of various kinds) but *not,* as we just said, those with which the *materialist* identifies such states. This suggests that psychophysical identifications (e.g., "pain is c-fiber stimulation") are not comparable to typical scientific identifications (e.g., "water is H_2O"); and this makes it questionable whether psychophysical identifications are correct.

Also, the fact that you could coherently imagine seeing water molecules (even though, for many reasons, our ability to see such things is limited) seems to be bound up (in a way that isn't easy to articulate) with the fact that identifications such as "water is H_2O" are *explanatory.* The identification of water with H_2O renders explicable otherwise inexplicable facts about water (e.g., that the amount of space occupied by a frozen gram of water is larger than that occupied by a liquid gram of water). But, at an intuitive level at least, it doesn't seem that a comparable gain of insight into the nature of pain is gained as a result of the identification of pain with c-fiber stimulation. Imagine somebody who, because of some defect, didn't feel pain. If that person were to learn everything there was to know about structure of the neural events underlying pain, he would still be ignorant of *what pain was.* [210] By the same token, somebody who has no idea what a neuron is can know exactly what pain is.

This suggests that psychophysical identifications are, in a fundamental way, not comparable to garden-variety scientific identifications. Somebody who knows what water tastes like, looks like, etc., doesn't necessarily know what water is. But somebody who knows that water consists of H_2O molecules really does know what water is—it doesn't really matter what other knowledge he has or lacks. It thus becomes very hard to sustain the idea that (A) "pain is c-fibre stimulation" is in the same category as (B) "water is H_2O." There is a clear sense in which somebody who knows (A) may not know what pain is but in which somebody who knows (B) cannot possibly *fail* to know what water is. In connection with this, there is reason to think that there is an *explanatory gap* between the mental and the physical—to believe, in other words, that there is no way to understand the mental in terms

of the physical. And if this is the case, then it would seem to follow that the mental is *not* the physical.

This argument evaluated

Even if the mental cannot be understood in terms of the physical, it doesn't follow that the mental isn't *identical* with the physical. Even if the fact that Jerry is sad cannot be understood in terms of the fact that Jerry's brain is in such and such a condition, it doesn't follow that Jerry's sadness isn't identical with his brain's being in that condition.

We know that biological phenomena (e.g., digestion, respiration) are identical with interactions of atomic and sub-atomic parts. But there would be no way to translate a biological statement (e.g. "Smith is now breathing") into a statement about atoms bouncing off of one another. Because statements belonging to biology involve concepts (e.g., respiration) that aren't found in physics, it is to be expected that the former are not going to be capable of being translated into, or otherwise understood in terms of, the latter. Biology talk isn't physics talk, even though all biological *events* are physical events. Similarly, psychology talk isn't physics talk (or chemistry talk or biology talk), even though psychological *events* are physical events. Because psychology uses concepts not found in any physical science (e.g., chemistry or physics), it is to be expected that statements belonging to the former are not going to be capable of being translated into, or otherwise understood in terms of, statements belonging to the former to the latter.[211]

Counter-response

Although this argument obviously has merit, and is widely accepted, I am not convinced by it. (But it is an argument that must be thoroughly understood.) There seems to me to be a significant difference between the relationship that holds between statements belonging to physics and statements belonging to biology, on the one hand, and the relationship holding between statements belonging to psychology and statements belonging to physics (or chemistry or biology . . .) on the other. This viewpoint is defended in Sections 9.1–9.2.1 of the present chapter.[212]

A counter-response to the counter-response

The mental *can* be understood in terms of the physical. There is no more of an explanatory gap between the mental and the physical than between the biological and the chemical. In other words, there is no explanatory gap. Right now we'll discuss why, according to many, there is no such gap.

9.1 Is there an explanatory gap?

So far as advocates of the view just described provide clear grounds for it, those grounds lie in two doctrines. Right now, we will discuss just one of them. (In a moment, we'll discuss the other.)

Some—for example, Carl Hempel (1905–1997)—hold that to explain why one state of affairs accompanies or follows another simply to show that , taken together, they instantiate a general regularity.[213] To explain why, in this one instance, smoke follows fire is simply to identify some comprehensive regularity of which this particular sequence of events is an instance. If it is shown that fire is *always* followed by smoke, it has *ipso facto* been explained why, in *this* case, fire was followed by smoke. Similarly, were we to establish perfect *concomitances* between mental events and physical events, we would *ipso facto* have explained the mental in terms of the physical.

A counter-response to the argument given

Hempel's conception of explanation is wrong. The basic problem is that it's too liberal: if it were right, anything could be said to be the cause of anything. The reasons for this are given in Chapter 17. But here's the basic idea.

I grab my phone, so as to pick it up and make a call; just after I do so, it rings. By coincidence, I'm getting a call at that exact moment. Hempel's theory has the obviously false consequence that my grabbing the phone is what caused it to ring. Let E_1 be the event consisting of my grabbing the phone, and let E_2 be the event consisting of the phone's ringing. Obviously E_1 will be unique in many respects—no other event will involve *exactly* the same mass-energy distributions that it does. It follows that E_1 uniquely has some property P_1. The same thing *mutatis mutandis* is true of E_2; and E_2

therefore uniquely has some property P_2. Thus, the sequence of events consisting of E_1 and E_1 is a unique instance of the true universal generalization that any instance of P_1 is immediately followed by, and adjacent to, an instance of P_2. If Hempel's analysis were right, then my grabbing the phone *would* have caused it to ring. Since it didn't, Hempel's analysis is wrong.

Another counter-response

Setting this aside, it isn't widely held that mind-brain concomitances (other than obviously contrived, explanatorily irrelevant ones similar to the one just described in connection with E_1 and E_1) are to be found. The reason: it is widely believed that the mental is "multiply realizable." Instances of any kind of mental category can, it is alleged, be constituted by physical events of very different kinds; and for this reason, there are no true statements of the form "events having such and such biochemical properties occur iff events having thus and such psychological properties occur."

9.1.1 How the functionalist tries to eliminate the explanatory gap

But, even though, for the reason just given, the alleged multiple realizability of mental states seems to eliminate one way of closing the explanatory gap between mind and matter, it actually opens what many (myself not included) claim to be a promising way of closing that gap. As previously noted, almost all contemporary materialists hold that mental entities *are* multiply realizable —that, for example, a belief that snow is white could be composed of virtual material. For X to be such a belief, it is said, is not for it to be composed of this or that material, but is for it to have certain *cause and effects.* They hold, in general, that given any mental category M, for a given physical entity P to fall into M simply *is* for P to have certain causes and certain effects. This doctrine is known as ***functionalism.***

In Sections 12.0–12.1.3 of the present chapter, it will be argued that functionalism is false, the reason being that it collapses into the obviously false view that there are *no* mental entities. It goes without saying that, if it's

false, functionalism does nothing to close the explanatory gap. In any case, I'd now like to put forth a positive reason to believe that there is no way to close the explanatory gap.

9.2 The reason there is an explanatory gap

For argument's sake, suppose that every mental event is identical with (or realized by) some physical event. In that case, my feeling remorseful that I stole my sister's cookie is identical with some brain-event of mine. Let M be my feeling of remorse, and let B be the corresponding brain-event. Suppose that I sincerely and accurately express my remorse—taking care to discuss its origins, its exact content, etc. Let SM be this statement. And suppose that I see an accurate image of B on a giant monitor, and I thoroughly and accurately describe what I see. Let SB be *this* statement.

SM describes the very same *event* as SB. But the *data* of which SM is a synthesis doesn't even overlap with the data of which SB is a synthesis. Even if mental *events* are identical with physical *events,* the *data* of which mentalistic *statements* are syntheses doesn't even overlap with the *data* of which physicalistic statements are syntheses. This means that, given the data of which any mentalistic statement is a correct synthesis, it is epistemically possible that none of the data of which any given physicalistic statement is a correct synthesis even exists. Mentalistic statements don't model the same data as physicalistic statements, or *vice versa.* So even though the *facts* described by the former probably are identical with (a subset of) the *facts* described by the latter, any statement belonging to the one category leaves it completely open whether any given statement belonging to the other is correct or not.

9.2.1 Davidson's argument for the explanatory gap

In his paper "Mental Events,"[214] Donald Davidson argues that there is an explanatory gap. Here is his argument:

> (DA[215]) Suppose the following. x is some event in Smith's brain. x also constitutes a desire on Smith's part to become a lawyer.
>
> Question: why does x constitute such a desire? Answer: because, given the thoughts and deeds on Smith's part that x follows and given the thoughts and Smith's part that x precedes, it makes *sense* to see x as a desire on Smith's part to become a lawyer. It is analytic that a desire on somebody's part to become a lawyer have certain causes and also certain effects. It is analytic that, given certain circumstances, a desire to become a lawyer lead one to study for the LSAT, apply to law school, and so on. And it's analytic that, given certain circumstances, a desire to become a lawyer result from one's belief that the lawyering is the best way to promote the interests of justice (or whatnot).
>
> Before we take the next step in our argument, we must generalize these points. Let P be an arbitrary physical event and let M be an arbitrary mental category. (An example of a mental category would be *belief that snow is white, desire to become a senator, regret at never having learned to play a musical instrument.* So for x to fall into the category *desire to become a senator* is for x to be a desire to become a senator.) If P is to fall into M, it is necessary and sufficient that, given what P's owner thought and did prior to P's occurrence, and given also what P's owner thought and did *after* its occurrence, it *makes sense* to suppose that P is an instance of M. In other words, for P to fall into M is for P to be such that it can be *reasonably interpreted* as falling into M. To be a belief that 1 + 1 = 2 is to be interpretable as being such a belief.
>
> This must be understood aright. Supposing that P is a belief that 1 + 1 = 2, there isn't anything *other* than P's being interpretable as being such a belief that makes it be such a belief. It's being thus

interpretable isn't a *consequence* of it's being such a belief: it *is* its being such a belief. In general, there isn't anything to anything's being an instance of mental category M other than it's being such that it can reasonably be seen as being such an instance.

To move forward, let us continue to suppose that P is a belief that 1 + 1 = 2. Let us also suppose that P is a brain-event and, in addition, that P falls into *physical* category C.

For P to fall into C isn't for P to be such that it can reasonably be interpreted as falling into C. It isn't in virtue of what P's owner does or thinks before or after P's occurrence that P falls into this as opposed to that physical category. If P falls into a given physical category, it is in virtue of its chemical structure, or some such, it being irrelevant whether so categorizing P has the consequence that P's owner is rational or not.

It follows that, even though mental events are identical with brain-events, the reasons for which a given event is mental are different from those in virtue of which it's physical. Moreover, the former reasons cannot be understood in terms of the latter. Thus, there is an explanatory gap.

Although I personally agree with Davidson that there is an explanatory gap, DA utterly fails to establish this. This is because it depends on two false assumptions.

The first is the assumption that to be a belief that 1 + 1 = 2 (or a desire to lose weight or an intention to buy a car, etc.) is to be interpretable as being such a belief (or desire or wish, etc.). In addition to being deeply counterintuitive, Davidson's "interpretivist" view of the mental is demonstrably false. According to Davidson's interpretivism, for brain-event x, had by person A, to fall into mental category M is for it to be the case that, if x is assumed to fall into M, then A is, to that that extent, rational. If this were right, nobody could possibly make any erroneous inferences; nobody could ever think or act irrationally. For, by Davidson's lights, if x doesn't have the consequences that a belief that 1 + 1 = 2 *should* have, it *ipso facto* isn't a belief that 1 + 1 = 2. But people make erroneous inferences from their beliefs; people aren't always rational. Therefore, Davidson is wrong.

Also, Davidson's argument assumes the truth of functionalism which, as previously stated, we will find to be false in Sections 12.0–12.1.3.

9.3 Descartes' arguments for dualism

Descartes provided a straightforward and, at an intuitive level, compelling argument for (1). Here it is. I can believe that x is a feeling (or belief or intention) without believing that x is a brain-state. Therefore, x cannot be a brain-state if x is a feeling (or belief or intention).

This argument evaluated

This argument is guilty of the so-called "intensional" fallacy. Remember what we said about intensional contexts: given only that "Smith thinks that the inventor of bifocals invented bifocals" is true, it doesn't follow that "Smith thinks that the first postmaster general invented bifocals" is true, *even though* the first postmaster general is identical with the inventor of bifocals. Similarly, given only that "Smith thinks that x is a belief" is true, it doesn't follow that "Smith thinks that x is a brain-state" is true, *even if* beliefs *are* in fact brain-states.

Let us put this more explicitly. The objects of doubt, belief, and all other propositional attitudes are (as you might guess) propositions. So when Descartes says:

(i) I can doubt the existence of a world that is not made of matter,

what he is really saying is:

(i*) I can doubt *that there exists a world that is made of matter.*

And when Descartes says:

(i) I cannot doubt the existence of my own mind,

he is really saying.

(i*) I cannot doubt *that my mind exists.*

The italicized propositions are distinct. So while it is true that they differ in their properties, since only one of them has the property of being capable of being doubted by Descartes, it doesn't follow that mind isn't matter, since the entities that, in this case, differ in their properties are not mind and matter, but are instead the two underlined propositions.

An aside: these reflections show that intensional fallacies arise, at least sometimes, when *grammatical* complements don't coincide with *logical* complements. The *grammatical* complement of the occurrence of "doubt" in (i) is "the existence of a world that is not made of matter," but its *logical* complement is "that there exists a world that is made of matter." The second expression denotes a *proposition;* the former does not. In general, intensional fallacies arise when it is assumed that, since the grammatical object of some expression doesn't denote a proposition, the *actual* object of that expression is also non-propositional.

A refinement of Descartes' arguments for dualism

Here is an improved version of Descartes' argument that, so its proponents believe, isn't vulnerable to the points just made.

I can *coherently* think:

(*) "x is a feeling, *but x is not a brain-state*."

In other words, (*) is not analytically false.

All necessary truth is analytic truth. In other words, if a statement is necessarily true, that is because it holds entirely in virtue of the concepts composing its meaning (i.e., it is analytic). Therefore, given that (*) is not analytically false, it follows that it isn't necessarily false and, therefore, that it *could* be true. So feelings *could* be distinct from brain-states.

Next-step: identity holds *necessarily* if it holds at all. If x and y are identical, then they are necessarily so. Proof[216]: suppose that x and y are identical, but only contingently so. Obviously x cannot be contingently identical with itself. So if x and y are contingently identical, then y has a property that x lacks: y has the property of being contingently identical with x, whereas x does not have the property. Thus, contrary to our supposition, x is *not* identical with y, showing that it is incoherent to suppose that there is such a thing as "contingent" identity.

Thus, given that (*) *could* be true, and given that identity holds necessarily if it holds at all, it follows that feelings (and, for exactly similar reasons, beliefs and pains) are *not* brain-states.

This argument evaluated

First of all, it is easy to adapt what we said about (i) and (ii) to this argument. So this refinement of Descartes' argument is, like the first, guilty of the intensional fallacy.

But there's another reason to question that argument. According to many contemporary philosophers, not all necessity is analytic. The sentence "water is identical with H_2O," it is said, expresses a proposition that, although non-analytic, is necessarily true. There is no possible circumstance where water is anything other than H_2O. There are circumstances where liquids that, in terms of their macroscopic properties, *resemble* water are not H_2O. But there are no circumstances where *water* is not H_2O.

Why is this? The categories in terms of which science explains the world are known as "natural kinds." Examples of natural kinds are *zebra, metal, proton, water, emotion, human being, acid, base, star* and *penguin.* Natural kinds are generated by properties that are predictively and explanatorily fecund. Properties that don't have those virtues (e.g., person who drives a green car, edible substance on Larry's table) don't generate natural kinds. The property of being H_2O generates a natural kind. Two things that are explanatorily and predictively very different can, at least within the narrow horizons within which non-experimentalists typically operate, be indistinguishable in respect of their macro-characteristics. Given that two things that are explanatorily and predictively very different can, for all practical intents and purposes, be indistinguishable in respect of their macro-characteristics it follows that the property of having such and such macro-characteristics does not determine a natural kind.

But two things that are indistinguishable in respect of their micro-characteristics (e.g., their chemical compositions) *cannot* be explanatorily and predictively different. Thus, the property of having such and such micro-characteristics generates a natural kind.[217] It can't be known a priori what chemical composition water has. But, for the reasons just given, its having that composition, whatever it turns out to be, is essential to it. Supposing that x is some body of water, x's being water is identical with x's having that composition. Since, therefore, water necessarily has that composition—and, therefore, is necessarily H_2O—and since it cannot be known a priori that it

has that composition, "water is H_2O" expresses a necessary, but non-analytic truth.

Evaluating this counter-response

This is a prima facie very compelling line of thought, and it is almost universally accepted. For reasons that I state in Chapters 8, 9, and 18, I do not accept it (even though I accept some parts of it; e.g., the part about how natural kinds are individuated by their microstructures or by whatever it is about them that makes them explanatorily and predictively useful). In any case, these reflections naturally bring us to Kripke's compelling argument for dualism.

2.4 Kripke's argument for dualism

All identities hold necessarily, and some identities are indeed a posteriori. Nonetheless, dualism can be established by means of an argument not entirely unlike Descartes'. Consider some (alleged) necessary, nonanalytic truth; for example:

(WH) water is H_2O.

(In this context, we'll assume for argument's sake that WH is non-analytic.) The reason that WH is non-analytic is that something that appeared, visually and otherwise, to be just like water might not *be* water, as it might have the wrong micro-structure. In general, whenever a statement of the form $x = y$ is non-analytic, it is because objects that aren't x can *appear* just like x. But it makes no sense to say that something that *appeared* (felt) just like a pain wasn't *actually* a pain. Anything that feels just like a pain *is* a pain. "Pain is identical with brain state X" would be non-analytic, if it were true. Thus, since what makes it possible for it to be non-analytic that water is H_2O is that *appearing* like water is different from *being* water, and the same thing *mutatis mutandis* being true of all non-analytic necessities, it follows that "pain is identical with brain state X" is not true.

10.0 Arguments for materialism

First argument for materialism: It is obvious the mental events create, and otherwise affect, physical events; and it is obvious that physical events create, and otherwise affect, mental events. A hot iron is pressed to your skin (physical event); as a result, you feel pain (mental event). You intend to go to the store (mental event); as a result, you end up going to the store (physical event). You feel nervous (mental event); as a result, your heart races (physical event).

There is extremely strong evidence to the effect that any physical event is the result *only* of other physical events. To explain why a person's body moves a certain way, it is necessary *only* to know what the strictly physical (neural, muscular, . . .) antecedents of that event are, and it is *not* necessary to know what its psychological antecedents are. The behavior of a person's body is no more in need of "mentalistic" explanation than is the behavior of a brick. (Of course, a knowledge of a person's psychological condition is often sufficient, given knowledge of easy-to-ascertain facts about the external world, for a knowledge of what that person is going to do next. But such knowledge is not *necessary:* such predictions can, in principle, be made without having any knowledge as to the relevant person's psychology.)

Given that every fact about a person's bodily behavior can be explained

in strictly physical terms, we can take one of two positions. We can say either:

(i) the mental does not, contrary to appearances, have any effects on the body,

or

(ii) the mental *does* have such effects—and the mental is *identical* with the *physical* events that we know to have those effects.

Since it is very hard to believe that the mental is without effect on the physical, (i) is almost certainly false. Hence (ii) is the right view.[218]

Evaluating this argument: My own personal view is that this argument is pretty solid, that being why I am a materialist. Others have responded (or, given their positions, would respond) differently to this argument.

For example, according to *epiphenomenalism,* even though the physical affects the mental, the mental does *not* affect the physical.

But epiphenomenalism is clearly wrong. It's an empirical datum that the mental affects the physical and *vice versa.* Should it be denied that this is a datum, or even that it's true, I would respond by showing that epiphenomenalism is at odds with a fundamental fact about the nature of causation. Causation is bidirectional. If my hand affects the table (e.g., if my hand pushes the table), then the table affects my hand. So if, as the epiphenomenalist believes, the physical affected the mental, then the mental would affect the physical. But in that case, epiphenomenalism is false.

The epiphenomenalist might counter-respond by saying that there are exceptions to the principle of bidirectionality that I've been advocating: physical situations create, and have effects on, shadows; but shadows don't have effects on anything.

But this move is a non-starter. Mental entities, unlike shadows, really do have causal powers. Even the epiphenomenalist doesn't deny that they affect one another. Given this fact, along with the fact mental events result from physical events, it follows that some of the energy associated with physical states of affairs is transferred to them. So unless energy is implausibly assumed to vanish from the physical world, never to return, we must assume that the mental can, so to speak, spit it back into the physical world and, therefore, that the mental *can* affect the physical.[219] And, as we've seen, once it is granted the mental can affect the physical, it must be granted that the mental *is* physical.

Another argument for materialism

As we've seen, the mental clearly affects the physical. If mental entities are subject to physical laws, they are *ipso facto* physical. (After all, what would it *be* to be physical, if not to obey physical laws?) And if something obeys physical laws, it is *ipso facto* physical. What would it mean to say "X obeys the laws of physics —but it isn't physical"? X is physical iff X obeys physical laws.

Thus, if mental entities are not physical, they are not subject to physical laws. So if mental entities were non-physical, things that *didn't* fall within the scope of physical law would constantly have effects on the physical world, resulting in physical events that couldn't be explained in terms of physical laws—in physical events that were, in fact, *violations* of physical law. But such violations do not occur. In any case, there is no evidence that they occur, and certainly no evidence that they occur with anything even approaching the frequency with which, if dualism were correct, they would occur.[220]

A counter-response

Surely we could imagine physical objects breaking physical laws. We can imagine people sprouting wings and flying—even though, given the structures of our bodies, such events would be violations of physical law.

A counter-response to the counter-response

Bertrand Russell said that objects *are* sequences of events that are governed by physical law. I agree with him.[221] If Russell is right, it makes no sense to speak of a physical object that is in violation of all physical laws. So when we imagine a person sprouting wings we *are* imagining something that is subject to physical laws (albeit laws that are different from those that are actually operative). Second, so far as the objects of one's thoughts *aren't* law-governed, those thoughts involve incoherent or otherwise impoverished views as to the nature of physical law. A defense of this controversial point of view lies outside the scope of the present book.

Third argument for materialism—Fodor's argument[222]

The mental obviously affects the physical. Anything that affects the physical is *ipso facto* physical itself. What is there to be physical other than being able to displace mass-energy (or whatever it is that events consist of)?

Evaluating this argument

I personally find this argument compelling, and I don't know of any cogent responses to it in the literature.

11.0 Interlude: the problems with dual aspect theory, eliminativism, and idealism

Supposing that the mental and the physical are two aspects of the same thing, we must ask: what is that thing?[223] It can't be mental; for surely the mental is not an *aspect* of the mental. (My belief that snow is white is not an *aspect* of itself or of any other mental thing.) And it can't be physical; for surely the physical is not an *aspect* of the mental. (The chair is not an *aspect* of itself or of any other physical thing.) So (3) is either incoherent *or* it presupposes that there is something spatiotemporal that is not either physical or mental. But it is incoherent to suppose that there is anything spatiotemporal that is not *either* mental *or* physical. It follows that (3) is simply incoherent.

It seems to me that so-called "dual aspect" theorists are really just materialists who have misstated their view. So far as that's correct, we can forget about (3), and focus on (2). And since we've already discussed (2), we

can move on to (4).

(4) is obviously a non-starter. But, for what extremely little it's worth, here's the argument behind it. Scientific hypotheses are supposed to deal with things that can be *publicly* verified. Suppose that scientist Smith insisted that he, and he alone, were able to see (or otherwise sense-perceive) the data that confirmed some hypothesis of his. Obviously Smith wouldn't be taken seriously. My pains, tickles, etc., cannot be witnessed by anyone other than myself. For this reason, some psychologists who were attempting to scientize their discipline said that such things had no place within science and, therefore, that they had no place in a scientific world-view. Given this position, it is a short step to saying that pains, tickles, etc., don't exist at all. In a word, science deals with what is objective; mental states are subjective; therefore, mental states don't exist.

There are many fallacies in this argument. One of them is that, in actuality, the existence of pains, tickles, etc., *is* easily verified. Of course, *you* cannot have *my* pains. But you can have your pains, and you can therefore confirm that pains exist.

Also, that argument involves a failure to distinguish two meanings of the term "subjective."[224] Sometimes when we describe something as subjective," we mean that it is influenced by personal biases. (*So and so's reaction to the incident was extremely subjective—it wasn't impartial and logical.*) But when we say that mental entities are "subjective," we mean—not that they are biased—but that they must be *had* in order to exist. There can't be orphan pains and beliefs floating around: a belief must be *somebody's* belief. Mental states thus have, as John Searle puts it, a "first-person ontology." Obviously science mustn't be "subjective" (i.e., it mustn't be biased); but it doesn't follow that it mustn't deal with the subjective (i.e., with things having a first-person ontology). In any case, it's a plain fact that mental entities exist, and eliminativism is a hideous monstrosity to which we shall give no further thought.

(5) is not a view that anyone really holds (even though a few—e.g., George Berkeley, William James, Bertrand Russell—have claimed to do so). (5) is really more of an epistemological view than it is a view about the relation between body and mind: so far philosophers advocate (5), it is on the

grounds that it is the only viable response to *skepticism about the external world.* (Skepticism is the focus of Chapter 12 in its entirety and of Sections 6.0–8.0 of Chapter 13.)

12.0 The different versions of materialism-behaviorism

There are different versions of materialism. Let us now consider them, starting with ***behaviorism.***

According to the behaviorist, to believe that 1 + 1 = 2 is only to have a disposition to respond with certain overt behaviors to certain physical stimuli. If when asked "what is 1 + 1?," you say "2"—if, in general, you respond to sensory inputs in ways that are characteristic of a belief that 1 + 1 = 2—then *in virtue of that very fact,* you believe that 1 + 1 = 2. In other words, according to the behaviorist, behaving in the right way *is* believing that 1 + 1 = 2. It isn't a *consequence* of such knowledge. It *is* such knowledge.[225]

The general definition of behaviorism is this. If M is any mental category, P is an instance of M just in case P is a disposition on the part of some individual to respond with certain overt behaviors to certain sensory inputs.

Behaviorism is false. How a person reacts to a given sensory input is a function, not only of what that input is, but also of what that person's existing mental states are. If you don't speak English, an utterance of "what is 1 + 1?" will *not* cause you to say "2." It is only in people who have various pieces of background knowledge (e.g., knowledge of linguistic conventions) that such a disposition corresponds to a knowledge that 1 + 1 = 2. What we said about that particular disposition is true (*mutatis mutandis*) of any other. By itself, no disposition to engage in a certain kind of overt behavior, given certain stimuli, corresponds to any mental state: what does so is such a disposition *plus* various pre-existing mental states. So mental states cannot be reduced to behavioral dispositions. At most, they can be reduced to behavioral dispositions *plus* various mental states. But this means that they *can't* be reduced to behavioral dispositions.

Aware of these problems with behaviorism, but still wanting to holding onto the idea that mental entities could be dispositionalized (identified with

tendencies to respond with certain overt behaviors to certain inputs), many philosophers of mind responded to this by adopting *functionalism,* which we've already discussed and will now discuss further.

12.1 Functionalism (revisited)

Consider the kinds of situations that characteristically give rise to a belief that snow is white (e.g., visual perceptions of white snow), and consider all the thoughts and behaviors that characteristically result from such a belief (e.g., one's forming the belief that snow is not green). According to the functionalist, for a brain-state to be a belief that snow is white just *is* for it to have the right causes and effects. So if brain-state B has the right causes and effects, B is *ipso facto* a belief that snow is white (or a desire to become president or a pain or a feeling of euphoria). B's having those causes and effects isn't *evidence* of its being such a belief; nor is it a *consequence* of its being such a belief. It *is* its being such a belief.

The functionalist says that, for any mental type M, mental entities are *among the characteristic causes and effects of instances of that type.* So the functionalist—unlike the behaviorist—allows that among the causes of (e.g.) a belief that snow is white are *other mental entities* (e.g., perceptions of white snow), and that among the effects of such a belief are other mental entities (e.g., a belief that snow is not green).

Remember that behaviorism falsely supposes that a person who had no linguistic knowledge would react by saying "2" to an utterance of "what is 1 + 1?" and, in general, that behaviorism identifies mental states with dispositions that can be understood in *strictly* behavioral terms. Functionalism doesn't run into this problem. Unlike behaviorism, functionalism can (supposedly) accommodate the fact that among the causes and effects of a mental entity are other mental entities. (Why "supposedly"? Keep reading.)

For something is (e.g.) a belief that snow is white wholly in virtue of its causal role. *In general, for any mental category M, there is some causal role C such that x falls into M if and only if x has C.* (This statement may not be easy to follow, but there is simpler ways of saying what functionalism is.)

12.1.1 Problems with functionalism

For something to be a pain is for it to *feel* a certain way—its causal role is secondary. It is true, of course, that pains are likely to have certain causes (kicks to the shins) and certain effects (screaming loudly). But for something to be a pain isn't for it to have a certain causal role; for something to be a pain is for it to feel a certain way.

Since it says otherwise, functionalism is wrong.

There is a related point. Supposing that P is some pain, P's having a given causal role is a *consequence* of its being a pain; its being a pain is not a consequence of its having that causal role. Supposing that P is some pain of yours, your writhing and screaming is an *effect* of P's being a pain and, contrary to what the functionalist alleges, it thus cannot be constitutive of its being a pain.

Let's move on. Functionalism says that x is a belief that snow is white iff x has the right causal role. But that causal role, we have seen, involves other mental entities. So it sounds as though functionalism analyzes the mental into causal role *plus* the mental: the mental is thus not eliminated—it's still there, making it questionable whether functionalism really is a form of materialism.

If cogent, this criticism of functionalism is devastating, given that the whole raison d'être for that doctrine is that it supposedly explains the relationship between mind and body. Functionalists respond by saying that, through judicious use of a technical device known as the "Ramsey sentence," it can be shown that functionalism *does* reduce the mental to the physical, while at the same time doing justice to the fact that, given any arbitrary mental type M, instances of M may have other mental entities for their causes and effects. Here is their reasoning:

> (RS[226]) We must distinguish the concept of *all* from the concept of *any*. Among the causes and/or effects of practically every given mental entity are (a) other mental entities and (b) strictly physical inputs (e.g., photons striking one's retinas) and strictly behavioral outputs (e.g., one's reaching for the glass). But even though *any given* mental entity is to be understood in terms of other mental entities, nonetheless the *totality* of a person's mental entities can be

> understood entirely in terms of strictly physical inputs and strictly behavioral outputs. Let MS (short for "mental sentence") be a single, very large sentence that says, for any given mental kind (e.g., *belief that snow is white, dread of visiting one's mother in law,* etc.), what the causes and effects of instances of that kind are. MS is going to contain "constants": it will contain expressions, such as "dread of going jogging in the morning," that refer to kinds of mental entities. Given any such expression, replace it with a variable. The result will be an *open sentence* (an expression that contains a *free variable but is otherwise just like a sentence*). *Bind* the occurrences of that variable in that open sentence with an existential quantifier. Do the same thing *mutatis mutandis* for each occurrence of any expression that refers to a mental kind. But don't do it for any of the expressions that refer to physical kinds. The result will be a sentence that completely describes the causes and effects of instances of each mental kind, but that does so without using any expressions that refer to anything mental. The mental will thus be completely *described* in terms of the physical. But that sentence will do justice to the fact that mental events cause other mental events to occur, and it will do justice to the precise way in which they do so.

By saying that causal role fixes content, RS makes content too brittle: *any* deviations of *any* kind between *any* two mental states *ipso facto* amount to differences in content. No two mental entities have *exactly* the same causal roles. (There isn't anything in your mind that has *exactly* the same causal role as my belief that 2 is an even prime. There isn't anything in any person's mind that has exactly the same causal role as anything in anyone else's mind.) It follows that, if RS is right, no two people can believe that 1 + 1 = 2 or that Mongolia is in Asia, etc. Since that is obviously false, so is RS.

A story may clarify this point. Both Smith and Jones believe that snow is white. Smith is a physicist who focuses on the reflective properties of white surfaces. Jones doesn't know anything about physics. Obviously the functional role of Smith's belief that snow is white will be very different from that of Jones' corresponding belief. But if, as functionalism holds, a brain-state's being a belief that snow is white just *is* its having a certain functional role, then presumably one person's belief that snow is white

shouldn't be able to differ, at least not *too* drastically, from another person's belief that snow is white.

12.1.3 The real problem with functionalism

It is not objects, but *states* of affairs, that have causal properties. What shatters the window is not *the rock*—it is rather *the rock's moving with a certain velocity, in a certain direction,* etc. Similarly, if B is the brain-state that realizes your belief that (e.g.) water quenches thirst, it isn't *B* that causes things to happen—it is B's having these or those properties. According to functionalism, B's being identical with a belief that water quenches thirst is *identical* with B's having certain causes and effects—it is identical with (e.g.) its causing you to drink water when thirsty. But if B's being identical with that belief is *identical* with B's causing you to drink water (when thirsty), then B's being such a belief cannot *cause* such behavior. Cause and effect must be distinct: if x causes y, then x is not identical with y (and neither is a part of the other). So if functionalism is correct, then *the fact that you believe that water quenches thirst* cannot cause you to do things like drink water when thirsty. Thus, if functionalism is right, then your *belief* that water quenches thirst does nothing.

Given any other behavior that we know to be caused by the existence of a certain kind of mental entity, an analogous argument shows that, if functionalism is right, such an entity cannot possibly cause that behavior. If a speaker of English who wishes to be sincere says "I believe that water quenches thirst," his uttering those words is caused by (among other things) his belief that snow is white. Functionalism says that there is some brain-state B such that that B's being a belief that snow is white is *identical* with (among other things) B's causing that person to utter those words. But if B's being such a belief is *identical* with its causing that utterance, then B's being such a belief cannot *cause* that utterance to occur. So functionalism *strips* the mental of any causal powers. Given that the mental *does* have causal powers, it follows that functionalism is wrong.

13.0 Conclusion

Given the fact that mind and body interact, along with the fact that such interactions seem not to involve violations of physical law, it seems that

materialism is probably correct. But there are two facts that, when taken together, pose a serious threat to materialism:

1. The mental *seems* to be different from the physical.

2. The appearance-reality dichotomy seems inapplicable to at least some mental entities (e.g., mental images), the reason being that some mental entities *are* appearances.

Taken together, 1 and 2 suggest that the apparent differences between the mental and the physical are at least sometimes actual differences.

Chapter 12

Skepticism and the Justification of Inductive Inference

1.0 What is skepticism?

A skeptic is a doubter, not a disbeliever. For me to be skeptical about Smith's assertion that he's an astronaut is for me to doubt that he's an astronaut. It isn't for me to be to *dis*believe that assertion—only to question it. If I'm convinced that Smith is not an astronaut, I'm not *skeptical* about his claim. Skepticism is doubt, not disbelief.

But there is some tendency for skepticism to give way to disbelief. Even if I never definitively learn that Smith is *not* an astronaut, if I continue to find no good reason to believe that he is one, I'm more likely than not to disbelieve it. We feel that, if a claim is true, some positive evidence for it will turn up: the mere absence of counterevidence isn't enough for us.

Does our human tendency to transition from doubt to disbelief have any logical basis? Or is it just another case of emotion and instinct corrupting rationality? It depends on the case. In some cases, but not all, failure to find positive evidence is itself counterevidence; and, in such cases, our tendency to slide form doubt to disbelief does have a logical basis. But, as we'll see in this very chapter, unless some highly defensible forms of skepticism can be refuted, this claim of mine is simply wrong.

Many important debates in philosophy concern the merits of some form of skepticism. Are there non-spatiotemporal entities? (Skepticism about platonic entities.) Do we have beliefs, feelings, etc., that we aren't aware of having? (Skepticism about unconscious ideation.) Could observation ever provide adequate grounds for believing in unobservables (e.g., electrons) or should we regard terms like "electron" as referring to fictions of some kind or as not referring to anything at all? (Skepticism about theoretical entities.)

As these examples indicate, philosophical skepticisms involve whole *categories* of truths. One doubts *all* truths affirming, or presupposing, the existence of non-spatiotemporal entities, unconscious emotions, theoretical entities, etc.

In this chapter, we'll discuss the two most important forms of skepticism: *skepticism about the external world* and *skepticism about inductive inference.*

To be a skeptic about the external world is to doubt that what our senses are telling us is accurate. One is a skeptic of this kind if one believes there to be no good reason to believe that one's sense-perceptions are anything other than hallucinations.

A bit of background is needed to say what it is for one to be a skeptic about inductive inference. Some knowledge is direct. One doesn't have to figure it out—one just knows it. Somebody who is experiencing a searing pain knows it, and he didn't have to figure it out. But much knowledge is indirect. Einstein didn't "just know" that a body's mass increases as its speed increases; he had to figure it out. In other words, he had to *infer* it. So for knowledge to be indirect is for it to be acquired through inference.

There are two parts to any inference: the premises and the conclusion. The conclusion is the belief arrived at—the thing that, when the inference leads from the right premises to the right conclusion in the right way, is "figured out." The premises are the pre-existing beliefs or assumptions on the basis of which one arrives at the conclusion.

There are two kinds of inferences: deductive and inductive. An inference is deductive if, supposing that one's premises are true and one's reasoning non-faulty, it is impossible for one's conclusion to be false. If my premise is that x is a square and my conclusion is that x has four sides, it's logically impossible for my conclusion to be false if my premise is true. So this would be a case of a deductive inference. An inference is inductive if, supposing

that one's premises are true and one's reasoning non-faulty, it *is* possible for one's conclusion to be false. If my premise is that, in the 35 years I've been drinking tap water, I have never become sick as a result of doing so, and my conclusion is that I won't become sick as a result of drinking the glass of tap water in front of me, my conclusion *may* be false. So this would be a case of an inductive inference.

A skeptic about inductive inference holds that it is impossible, as a matter of principle, to learn anything through non-deductive inference. He holds, in other words, that one cannot acquire knowledge through inductive inference. All knowledge is either direct or is arrived at through deductive inference.

If this is right, then we know only an infinitesimally small fraction of what we think we know. This is because, for reasons thoroughly discussed in Chapter 10, no knowledge about the external world is direct, and only some knowledge of ourselves is direct.

We will start by discussing skepticism about induction.

2.0 Skepticism about induction

David Hume made it clear why it may be doubted whether inductive inference ever leads to knowledge. Here is Hume's argument:

> (HA[227]) Consider some inductive inference—for example, "the sun has always risen in the past; therefore, it will rise tomorrow." First of all, this argument is not *deductive,* since, even if the premise is granted, it isn't a theoretical impossibility that the conclusion will be false. The sun might explode, after all. So the conclusion is, at best, probable. That's step 1.
>
> Here's step 2. Given that the connection between premise and conclusion isn't deductive, we must ask: *how* does the premise warrant the conclusion? In other words, what reason does the (supposed) truth of the premise give us for accepting the conclusion?
>
> Here's an answer that seems tempting at first: "The past resembles the future; the known resembles the unknown. Nature is uniform. So given that such and such has been the case until now, we may infer that such and such will continue to be the case. Given that, until now,

pinewood has floated, metal has expanded when heated, water has quenched thirst, etc., it's reasonable to conclude that pinewood will continue to float, and so on. It is thus the principle that nature is uniform—or, as we'll call it, the *uniformity principle* (UP)—that allows us to make rational, non-deductive inferences about the unknown on the basis of the known."

But this argument is no good. No one directly knows that UP holds for all places and times. At most, one knows that it *has* held for the fragments of nature that one has personally known of. One doesn't directly know that it held of past objects that one didn't know of; and one doesn't directly know that it will hold of future objects that one will know or of future objects that one won't know of. Any such knowledge must be indirect (i.e., must be inferential). The inference in question can't be deductive. Given only that the ravens thus far known to one have been black, it doesn't follow *deductively* that ravens not (yet) known to one are also black. In general, given only that such and such has held of the things I've known of, it obviously doesn't *deductively* follow that it holds of the things that I haven't (yet) known of. The inference must be inductive. But if it's inductive, then it goes through only on the condition that UP is granted. But since that very inference must go through *in order* for UP to be granted, it doesn't go through and, what is more, UP cannot be granted and, finally, there's no way to justify inductive inference.[228]

3.0 Problems with Hume's argument

Nobody denies that inductive inferences are knowledge-conducive and, therefore, that the conclusion of Hume's argument is false. Even Hume said that, in his bones, he couldn't accept it. But there's no accepted response to Hume's extremely clever argument. Nonetheless, Hume's argument is refutable and a refutation of it must lie in some argument not totally unlike the following.

There are two serious errors in Hume's argument. First, it assumes that the primary form of non-deductive inference is *induction by enumeration.* In actuality, induction by enumeration is less important than, and subordinate to,

a very different form of non-deductive inference known as *inference to the best explanation.* (The italicized terms will be defined shortly.) Second, Hume assumes that non-deductive inferences go through only if UP is granted (i.e. only if it's granted that nature is uniform.) But UP is not the relevant principle. What must be granted, I propose, is that good explanations minimize causal anomalies. What must be granted, in other words, is that, other things being equal, theory T_1 is better than theory T_1 if the world as T_1 describes it contains fewer unexplained causal connections—fewer causal links that must be taken for granted and left unexplained—than T_1.[229] Let's refer to this principle as "MC" (short for "minimization of causal anomalies").

Unlike UP, MC is, I submit, *analytically* true: it is inherent in the concept of what an explanation is that, other things being equal, the theory that eliminates the greater number of unexplained explainers—and, in particular, the greater number of unexplained causal connections—is the one to be preferred. Because this principle, unlike UP, is analytically true, it doesn't have to be validated on inductive grounds. This makes it possible to validate inductive inference in a way that, unlike the way that *Hume* considers, is not viciously circular.

3.1 Inference to best explanation versus enumerative induction

There are two forms of inductive inference: induction by enumeration and inference to the best explanation. (See Chapter 12.) An induction by enumeration has the form: every F that I've ever known of was also a G; therefore all F's are G's. Example: All the balls that I've thus far removed from the urn have been white; therefore, all the balls in (or removed by me) the urn are white.

An inference to the best explanation has the form: it is known that S; it isn't yet known that S*; but if S* were true, it would, without doing excessive violence to what we (think we) know about the world, explain why S is true; moreover, if S* weren't true, it would be hard to find a viable alternative explanation of S. For example: I'm very careful with my credit cards (I never use them to make online purchases, I never let anyone besides

reliable merchants see them, etc); and, before my friend Larry moved in with me, no unauthorized purchases were ever made with any card of mine. But after he moved in, my cards were used to make a series of unauthorized purchases. Moreover, the items bought were ones that, for work-related reasons, Larry desperately needs and that nobody outside of Larry's rather unusual profession would have slightest use for. Knowing all of this, and not knowing any other way to account for the unauthorized purchases, I conclude that Larry made those purchases.

It's widely assumed among philosophers that induction by enumeration is the more fundamental form of inference. This is a mistake. All cases of induction by inference, when legitimate, are parasitic on cases of inference to the best explanation. There are, of course, many cases where *one* of my reasons for thinking that all F's are G's is that every F I've ever known of was also a G. But, setting aside beliefs arrived at through spurious reasoning, there are no cases where that was my *only* reason for believing that all F's are G's. Hume's argument focuses entirely on induction by enumeration. Hume is right, I believe, that *this* form of inference is unjustifiable. But we don't use this form of inference. In any case, we don't use it in the way that Hume thinks. Moreover, we shouldn't use it, and we needn't use it. In some cases we may *think* that we're doing straight induction by enumeration. But, in almost all such cases, a little probing always shows that many interconnected background assumptions are at work and that, when these are made explicit, what we're really trying to do is to produce an inference to the best explanation. And, although there are cases of straight induction by enumeration, they tend to occur in rather artificial contexts and that, when they do occur, they're *clearly* spurious.

There is a well-known inductive error known as the "gambler's fallacy." (Sometimes it's referred to as the "Monte Carlo" fallacy.) You are playing a game of roulette. (This game involves rolling two dice into a spinning wheel, which is divided into numbered slots.) You know that nothing about the game is in any way rigged—that the dice aren't loaded, and that neither the wheel nor the underlying apparatus were built or modified to favor any one outcome more than any other. And you thus know that, given only the physical properties of the game, a die, once thrown, is no more likely to land in this as opposed to that slot. (How you know this is irrelevant; we'll just assume that

you do.) You *also* know that, given any slot on the wheel, there is nothing about you that makes a die you've thrown more likely to land in that spot than in any other—nothing about your body, mind, or dice-throwing practices that favors this as opposed to that outcome. (Again, *how* you know this is irrelevant; we'll just assume that you do.) You've rolled the dice seven times. Each time, at least one of the die has landed in slot #27. It must be stressed that you know there to be nothing about the game or about you that predisposed this to happen—that you know this to be entirely a matter of luck.

Under these circumstances, it would be natural to predict that you'll roll a 27 next time. But that prediction would be irrational: you have no more reason to believe that you'll roll a 27 than you will any other number. And if, by chance, you *do* roll a 27 next time, it will be natural to predict that you'll roll a 27 the *following* time. But this too would be completely irrational. So long as you know that there's nothing inherent in the game or in yourself that favors this as opposed to that outcome, you are never any more entitled to bet that, with the next throw, you'll roll a 27 than you are to bet that you'll throw any other available number. It doesn't matter how many consecutive 27s you rolled prior to that throw: it could have been ten, a thousand, or a million.

By the same token, no matter how many consecutive 27s you roll, you are never any more entitled to bet that, next time, you *won't* roll a 27 than you are to bet that you won't roll any one of the other available numbers. It would be natural to think otherwise, of course. Given how improbable it is that you will roll five consecutive 27s, let alone 50, you may have an urge, when rolling for the 51st time, to think: "*this time,* surely, I'm even *more* unlikely to roll a 27 than I am to roll any other available number. The game isn't rigged. The likelihood of my rolling 51 consecutive 27s is infinitesimal and, therefore, is much smaller than 1/n (where n is the number of available slots). So a 27-roll is *especially* unlikely—more unlikely than a 5-roll or a 13-roll." But this is spurious reasoning. You are no more, and no less, *un*likely to roll a 27 than you are to roll any other number. With every throw, the chances of rolling any one of the available numbers are identical with the chances of throwing any one of the others. Since, by hypothesis, the conditions that obtain on any two occasions are identical, it would be irrational on any given occasion to think any one outcome any more or any less likely than any other.

It would, in fact, be the ultimate absurdity—like the belief that, although doughnut x is an atom-for-atom duplicate of doughnut y, x is far more fattening than y.

3.1.1 When is enumerative induction rational?

When *would* it become rational to believe that, next time, you're more likely than not to roll this as opposed to that number—that, for example, you're especially likely to roll a 27? This belief becomes rational when, and *only* when, you have reason to believe that a 27-roll is favored by the structures involved in the game. And that belief, in its turn, is rational if you know that circumstances at all like the following obtain:

*The dice are magnetically attracted to the 27-slot.

*On any given occasion, you have an unconscious intention to roll a 27 (even though you have no conscious intention of doing this), and you're such a talented dice-thrower that, if you can roll a 27 if it is your (subconscious) intention to do so.

*The 27-slot is much bigger than any of the other slots. In fact, it takes up so much space on the roulette wheel that the remaining spaces are too small for the ball to fit into them.

You are rational to believe that you'll continue to roll 27s to the extent that your having thus far rolled multiple 27s in a row gives you reason to believe there to be some underlying structure favoring that outcome. And to the extent that a long run of 27-rolls *doesn't* give you such a reason, you are *irrational* to believe that you're any more (or any less) likely to roll a 27 than you are any other number. So, no matter how many consecutive 27s you roll, if you know with complete certainty that nothing about either yourself or the set-up of the game favors that outcome, it is irrational to think that, next time, you're any more (or less) likely to roll a 27 than you are a 5 or a 32. Put pedantically, it is only insofar as you have reason to believe in such a structure that you have reason to expect something that has the property of being a die thrown by you to have the property of landing in the 27-slot.

3.1.2 Generalizing these points

Your knowing of many phi's that are psi's and of none that are not doesn't

necessarily give you any reason to believe that the next phi you encounter will be a psi; it gives you such a reason *only* insofar as it gives you a reason to believe in some structure or mechanism that disposes phi's to be psi's. If you know on independent grounds that there is no such mechanism, no run of phi's that are psi's, no matter how long, gives you a reason to think that the next phi will be a psi; and you are guilty of the gambler's fallacy—and thus, as we saw, of the most rank absurdity—so far as you think otherwise. Thus, *in and of itself* induction by enumeration is worthless; it's a fallacious form of reasoning and falls in the same category as every other argumentative fallacy. Of course, if you know of many phi's that are psi's and of none that are not, you *do* have good reason to hold that, other things being equal, the next phi you encounter will be a psi. But this is because you're having this knowledge gives you reason to believe there to be a principled or lawful connection between a thing's being a phi, on the one hand, and its being a psi, on the other. Here, then, is the *real* structure of inferences that are wrongly thought of as cases if induction by enumeration: (i) you know of many phi's that are psi's and of no phi's that are non-psi's; (ii) on the basis of step (i), you believe there to be some principled connection between a thing's being a phi and its being a psi; (iii) if you know of a given thing that is a phi, but don't yet know whether or not it's a psi, you know, given (ii), that it's likely to be a psi; therefore, (iv) you conclude (if only tentatively) that it *is* a psi.

Thus, any case of induction by enumeration that isn't an instance of the gambler's fallacy involves the positing some mechanism or law that, were it to exist, would explain a certain concomitance—it involves, in other words, a case of *inference to the best explanation.* The best explanation of the fact that all known phi's are psi's is that, thanks to some mechanism or, in any case, principled connection of *some* kind or other, a thing's being a phi disposes it to be a psi; and as soon as such a mechanism has been posited, but not one second sooner, your knowing of a given thing that it's a phi is a reason for you to hold that it's a psi. So any non-fallacious "induction by enumeration" is nothing other than a case of inference to the best explanation (or, more precisely, is a case of applying some theory that embodies an inference to the best explanation).

Hume's argument assumes that it is only through induction by enumeration that the past is any guide to the future. It assumes that, so far as

we have any reason to believe that future phi's will be psi's, it is that past phi's have been psi's. But this assumption is dead wrong. The fact that past phi's were psi's, is not, in and of itself, reason to hold that future phi's will be psi's; it is such a reason *only* to the extent that it suggests some mechanism that disposes phi's to be psi's. Hume's argument assumes otherwise and is thus guilty of assuming the legitimacy of the spurious form of inference involved in cases of the gambler's fallacy. And, if it is to have a chance of going through, Hume's argument must be so reconstructed that it no longer embodies that assumption.

4.0 The minimization of causal anomalies

We haven't yet refuted Hume's argument—we've only taken the first step towards doing so. Hume could defend his view against what we've said thus by far by saying the following:

> (HR[230]) Suppose that, to explain why all phi's thus far known are psi's, you posit some underlying structure or law that disposes phi's to be psi's. Unless you think that nature is uniform, you have no right to expect that connection to *continue* to hold. But if, in order to deal with this, you suppose that nature is uniform, then you're caught in the vicious circle that I described.

HR is correct. One is indeed caught in a vicious circle if, in order to show the legitimacy of inductive inference, one assumes UP; and the reason is that, just as Hume says, UP can be known, if at all, only on inductive grounds.

4.1. Why inductive inferences don't presuppose the truth of the uniformity principle

But in making an inductive inference, one doesn't assume UP and, moreover, one doesn't assume *anything* that, like UP, can be known only on inductive grounds. What one assumes is that explanations are supposed to eliminate causal anomalies—that they are supposed to reduce the number of them and to limit the scope of those that aren't eliminated. What one assumes, then, is that it is inherent in the very concept of explanation that, other things being equal, T_1 is a better explanation than T_2 if T_1 does a better job of eliminating

causal anomalies than T_2.

The purpose of explanation is to minimize the breadth and depth of what must be taken for granted. The more a proposed theory requires you to say: "things just happen that way; there's no explaining it," the less successful an explanation it is.

Here's an illustration. On Monday night, you park your car in the usual place, viz. right in front of your house, which is in a quiet residential neighborhood. As usual, you make sure that you lock the car and turn the car alarm on. You also put an almost, but not quite, indestructible device (popularly known as "The Club") on the steering-wheel that locks it into place, making the car undriveable. Given where your home is in relation to where the car is parked, you'd almost certainly hear the car alarm go off, were it to do so. In fact, unless you were fast asleep, you'd even hear the footsteps of someone approaching the car. So even if the car alarm weren't on, you might well hear the approach of a would-be thief and you'd certainly hear the noises made by somebody who, not having the keys needed to open the car or to unlock the steering wheel or to start the car, would have to jimmy the door lock, saw through the wheel lock, etc.

As it happens, you hear nothing all night. And you know that you didn't sleep any more deeply than usual—that, for example, you weren't in a particularly deep drug-induced slumber—and, in general, that you weren't made less sensitive to noise than you usually are. (For argumentative purposes, we'll set aside the question of how you know this.) Nonetheless, on Tuesday morning, the car is gone.

What are the various possible explanations of this? (By "possible," I don't necessarily mean "worthy of consideration," only "not ruled out by the laws of logic.") There are infinitely many; but here are three.

(E_1) Using his or her super powers to bypass the security measures you took, some superhero-like creature made off with the car, making little or no noise and leaving few or no clues.

(E_2) As you were sleeping, the laws of physics changed in such a way that hitherto noisy processes (e.g., those that occur when car alarms go off) are no longer noisy.

(E_3) Your good friend Larry—who, despite his numerous convictions for auto-theft, simply radiates trustworthiness and decency, and to whom, without so much as a touch of anxiety, you therefore gave copies of all your keys—made off with the car. (For argumentative purposes, let's assume that, given that the car is gone, it had to have been stolen; it could not, we will suppose, have been towed or simply borrowed.)

Each of E_1 and E_2 replaces the mystery of how your car could be noiselessly stolen with a far greater mystery. In fact, each creates a whole network of deep mysteries. E_1, if correct, would generate a number of thorny questions concerning specific matters of fact and also concerning our knowledge of biological and physical law. (The features that we'd have to ascribe to the superhero in question would be ones that, unless modern biology is very wrong, a life form is most unlikely to have.) And E_2, if correct, would do nothing less than create the greatest scientific mystery of all time.

E_3 is not free from sin either. In general, you are a good judge of character. In any case, that's what you believe and that's what others tell you; and your judgments about people have generally been borne out by subsequent events. You've never been as convinced of anyone's integrity as you are of Larry's. Without being morose or censorious, he exudes a decency and centeredness that you've never even seen in a head of state, let alone a car thief. And although Larry's extensive criminal background is consistent with his stealing your car, it is not, you are convinced, *as* consistent with it as his current excellence of character is *in*consistent with it.

But, of course, E_3 is hands down the best explanation. You may think you're a good judge of character; but that doesn't mean you *are* one. People proverbially overestimate their ability to read others. And even if you are a good judge of character, so what? Anybody can be fooled; everybody *is* fooled, at some point or other. And if Larry is a career criminal—which, given his rap-sheet, he may well be—then he has probably honed his skills as a con-man.

This doesn't mean that E_3 is completely innocuous. If correct, it raises the question of how somebody could seem so different from how they are. It also raises the question of how, despite being a crime expert, you could have been unwary enough to let yourself be seduced into trusting somebody who, given his background, was so obviously unworthy of trust. But these mysteries can be dealt with more easily than the mysteries created by E_1–E_2.

All of this goes to show that *ceteris paribus* the best explanation is the one that leaves us with the fewest unexplained explainers.

Of course, new theories raise questions of their own. This is almost a tautology. A new theory *is,* tautologously, a new way of looking at things; and a new way of looking things is, very obviously if not tautologously, one that raises new questions. But good new theories tend to facilitate the answering of the questions they raise. Relativity Theory raises questions that don't arise for its predecessor, Newtonian mechanics.[231] But Relativity Theory itself tends to contain hints as to how those questions are to be answered. The development of a good theory tends to consist largely in making explicit what is already present in that theory. The development of a bad theory consists in *adding* new material to it to help it deal with each new onslaught of counterevidence.[232] In any case, what's important in this context is that good theories get rid of causal gaps and they shrink the ones they can't get rid of.

Before proceeding, we must make a certain distinction very clear. Little or nothing about the *actual* causal structure of the world is analytic. It isn't analytic that there aren't giant causal gaps in the causal structure of the world; that reality doesn't consist of one *ex nihilo* event after another, that superheroes don't suddenly pop into existence, steal cars, and then, without leaving a trace, vanish never to be heard from again. These things, if known, are known only on the basis of sense-experience. But it *is* analytic that, other things equal, good explanations are less dependent for their truth than bad ones on the existence of unexplained explainers. Given that E_3 depends less for its truth on unexplained explainers than does any one of E_1–E_2, it is *analytic* that E_3 is *ceteris paribus* the better explanation.

But it isn't analytic that E_3 is *correct.* It's a possibility—maybe only a bare theoretical possibility, but a possibility no less—that the laws of physics changed over night, that CIA super-agents have targeted you, that comic book characters now walk the Earth, etc. Anything's possible. And so far as we know that E_1–E_2 are wrong, it's through sense-perception. So, I repeat, it isn't analytic that E_3 is correct. But, given that it requires fewer unexplained explainers than others, it *is* analytic that *ceteris paribus* it's the better explanation.

4.2 Correct explanations can be bad explanations

For an explanation to be *good* isn't for it to be correct. Sometimes the right explanations are bad ones. A story will make this clear. I'm on a bus. The bus driver is smiling. A mystery! "What on Earth does he have to smile about?" I ask myself. His job is so boring, and his life must therefore be such a horror." But then I remember that, just a minute ago, a disembarking passenger gave him fifty $100 bills as a tip. So I have my explanation: "he just came into a lot of money."

But here is the very different explanation tendered by my seatmate Gus, who, in addition to being unintelligent, is also completely insane. "The bus-driver is a CIA assassin. This morning he killed somebody who, by coincidence, had the name 'Benjamin Franklin.' Benjamin Franklin (the statesman, not the murder victim) is on the $100 bill. So when the bus driver saw those bills, he immediately thought of that morning's murder. The murder was a particularly enjoyable one; the bus driver had a chance to try out some innovative techniques. So, on being given the fifty $100 bills, the bus driver was reminded of that happy experience." *That,* my seat-mate believes, is why the bus driver is smiling. The bus driver is indifferent to the fact that he is now richer by $5,000 than he was before. (He's not a materialistic man. Plus, the CIA pays him a huge salary. $5,000 is nothing to him.)

Gus and I have access to the same empirical data. (Gus hasn't read the bus driver's diary; he doesn't know the bus driver any more intimately than I do; and so on.) And Gus doesn't have some sort of psychic gift that would give him access to otherwise unknowable facts about the bus driver's mind

that would legitimize his explanation. Indeed, a belief that he has such a gift isn't even among Gus's many delusions. Nor does he have any good reason —even if, for argument's sake, one allows his delusions and hallucinations to count as good reasons—to believe that he has such a gift. So given the entirety of the data at our disposal, Gus's explanation is a bad one. The *only* evidence in favor of Gus's theory is that the bus driver started smiling upon being given the money.

But Gus is right. His explanation is correct down to the last detail.

Given that it turned out to be correct, should we say that, despite first appearances, Gus's explanation is not a bad one? No! It's a datum that it's bad. It's a bad explanation that turned out to be correct. Thus, for an explanation to be a good one isn't for it to be correct.

But a short extension of our story illustrates the even stronger principle that for an explanation to be a good one isn't even for it to be *likely* to be correct. Let W be the real world and let W* be a hypothetical world where we have the exact same experiences that we have in W, but in which the Gus's of the world are always right.[233] In W*, it's because Santa Claus is personally delivering gifts to millions of people that they suddenly appear in living rooms round the globe; it isn't because hard-working parents purchased them. In W*, your keys are missing because stealthy gnomes took them; it isn't because you drunkenly threw them in the incinerator. And so on. In W*, the right explanations are the ones that, in our world, are extremely bad ones. The data on the basis of which those hypotheses are held in W* (so far as those hypotheses are held) is identical with the data on the basis of which they're held here (same qualification). But in W*, they're right, and the good explanations—the logical, sane, well-supported explanations—that they compete with are always wrong. So, in W*, good explanations aren't even *likely* to lead to the truth. So there's no inherent connection between an explanation's being good and its being true: goodness (in the explanatory sense) cannot be identified with, or even understood in terms of, truth-conduciveness.[234]

Explanatory goodness is a relationship, not between explanations and reality, but between explanation and data. The relations that hold in W* between theory and data coincide with those holding in W. That's why the

ones that are wrong in W* are no less good on that account. What is it that, in W*, makes wrong explanations be good? In general, what is the relationship that good theories bear with respect to the data?

First of all, "good" is a relative to term. An explanation is good *relative* to some body of data. No good explanation is inconsistent with the relevant data. And, of the various explanations consistent with a given body of data, the best one is the one that depends the least on unexplained explainers.

Here's an illustration. I put an opaque plate on a table. I thus can't see the part of the table covered by the plate. "Given only this data," says my sophist friend Smith, "it's a theoretical possibility that the moment you put the plate on the table, there suddenly came into existence a hole in the table the exact size of the plate." Supposing that you're right, I ask Smith, why didn't the plate fall through the table? "Because," Smith responds "when that hole came into existence, a special force was created that kept the plate in exactly the place that it would have occupied if there were no hole." Why, I ask in response, did that force come into existence? "I don't know," says Smith. "It just did."

Maybe this explanation is correct. We'll never know for certain that it isn't—at least not without presupposing the falsehood of various other, comparably *ad hoc* explanations.

Be that as it may, it's a bad explanation. Why is it bad? Because it leaves us with more questions than answers. Why did the hole appear? Why did it happen to have *just* the right shape and size? Is it because the plate has special chemical properties? If so, why is it that, the moment the plate is whisked away from the table, the effects of any chemical interaction between the two are undone? Supposing that these questions can be answered, why did some force suddenly step into the breach left by the momentarily absent table surface? Why was that force *just* right—*just* enough to keep the plate on the table, but not so much that the plate rose above the table or trembled or did anything else inconsistent with its being on an ordinary table? And supposing that we have an answer to *that* question, why was there no independent evidence of the existence of that force? Why was its *only* effect to keep the plate in that very location, for that exact amount of time? Supposing that, for some reason or other, we have highly sensitive instruments of measurement underneath the table, why didn't any of them

detect anything?

With every answer, a new hole arises. It may be that, for each hole that pops up, there is a way to plug it up that is not only consistent with the laws of logic but is *also* consistent with our other empirical beliefs. Or, if there arises a case where the hole cannot be plugged up without jettisoning some, or even all, of our remaining beliefs, maybe there is a way to revise those other beliefs in such a way that they remain consistent with the raw data *and* with the stopgap in question. It is generally said that a given body of data can be modeled by many different, mutually inconsistent hypotheses.[235]

But even if this is true, it doesn't follow that, for each hypothesis that models a given body of data in a way that has a given degree of success in the way of eliminating unexplained explainers, there is a logically incompatible way of modeling the data that also has that degree of success in that respect.

4.2.1 Correct explanations can be bad explanations (continued)

A continuation of our plate story will make it clear what this means and why it is true. Why doesn't the plate fall through the table? Because, says the conventional wisdom, the table didn't vanish. But, the skeptic will retort, maybe we can jettison that explanation; maybe there's an alternative to it that is equally compatible with the perceptual data.

But, I would respond, supposing for argument's sake that there is such an alternative, it doesn't follow that it will create fewer holes to plug up than its predecessor did. It doesn't follow that it's comparable from a cost-benefit standpoint to the old one—that, in other words, it creates as few *new* mysteries as the old hypothesis while also eliminating as many existing ones.

But not only does this not follow: the exact opposite holds. Consider the alternative explanation that we considered a moment ago (viz. *when the plate is placed on the table, a hole opens up,* etc.) The explanatory cost of that hypothesis (in other words, what it leaves us having to explain that we didn't have to explain before) was a thousand times greater than its explanatory value (in other words, what no longer has to be explained that did have to be explained before).

It would be very hard to produce an explanation that was incompatible with the conventional one and, while remaining consistent with the relevant

data, was as inexpensive from an explanatory viewpoint as the conventional one. I invite you to verify this for yourself. If there is *any* reason to believe that there exists so much as a single such hypothesis, it isn't that such a hypothesis has ever been produced or ever will be. That reason, if it exists, must be purely a priori. And, for reasons that will presently become clear, there can be no a priori reason to accept such a hypothesis.

4.2.2 Correct explanations can be bad explanations (continued)

But there is an even more immediate problem for this (or, by an analogous argument, any other non-standard) substitute for the conventional view that the table, even when obscured by the plate, continues to support it. It's far from clear that, once *all* the relevant data is taken into account, that hypothesis is even *logically* possible. Let's look at the statement:

(*) "when the plate is placed on the table, a hole that coincides with the under-surface of the plate opens up. But a force field is suddenly created that keeps the plate exactly where it would have been if, other things being equal, that hole hadn't opened up."

Taken by itself, (*) describes a situation that could hold. That statement is a coherent one. It's not in the same category as "squares are round."

But, while that's obviously *necessary* for its being a logically feasible way of modeling the relevant data, it's far from sufficient. Think of the various physical laws that must be broken for the table section underneath the plate to vanish. Violations of natural laws, were they possible, would not be innocuous; there would be ripple effects. Think of the forces that (so we believe) *actually* prevent that plate-sized chunk of table from vanishing. Think of the various equilibriums that would be altered if those forces were immediately suspended. Think of the various energy transfers that would be changed or destroyed if, all of a sudden, those forces were put out of commission.

Those forces don't *only* keep that chunk of table in place. Were they all to be suspended, there would be many consequences *other* than the disappearance of the requisite table chunk. In other words, there would be a lot of explanatory "collateral damage," meaning that many events would occur *other* than the ones that, given that our objective is to validate (*), we

want to occur. And there's no guarantee that those other events would be consistent—even in the narrowest, most emptily logical sense—with the data that, supposedly, is modeled just as well by (*) as it is by the conventional explanation (viz. "the place doesn't fall because the table chunk doesn't disappear") that it replaces.

4.2.3 Correct explanations can be bad explanations (continued)

We might try to deal with this batch of problems by positing *second-order* suspensions of natural law. So in addition to suspending those laws involved in keeping the table chunk in place, we *also* suspend those laws that, given the first suspension, lead to the just-mentioned unwanted collateral events.

The problem is that those second-order suspensions of law will be no more innocuous than their first-order counterparts. To counteract the unwanted effects of suspending the first set of forces, we have to suspend another set of forces. But just as the original forces did a lot more than keep the table intact, so these new forces will do a lot more than suspend the unwanted effects of suspending the first set of forces. And it's obviously of no use to try to deal with this problem by bringing in a third set of forces.

"You may be right," it will be said, "that *if* we hold onto our existing beliefs about physical law, *then* the disappearance of the table chunk would have all of this unwanted fall-out. But surely we could adjust those beliefs in such a way as to prevent this. Whatever those beliefs are, just get rid of them and replace them with ones that *don't* lead to all of this unwanted fall-out."

The very fact that this move must be made only proves my point, namely, that (*) is *not* consistent with what we know. In fact, it proves a stronger point, namely: (*) isn't consistent with *any* body of beliefs *at all* like the ones we now have.

The objector is saying, rightly, that there is *no* way to square (*) with our existing beliefs—that it is only what we believe *minus many of our beliefs about natural law* that can be so squared. But when our beliefs about natural law are subtracted, so to speak, from the totality of our beliefs, the resulting body of beliefs bears very little resemblance to our current beliefs. Our beliefs about natural law are not in all cases neatly separable from our beliefs about specific matters of fact; and if we were to strip away enough of the former to compatibilize (*) with our belief system, we'd be biting off a lot

more than we bargained for. This is because various beliefs about natural law are embedded in the concepts associated with expressions that, in order to say what's on our minds, we must use every day (e.g., "microscope," "telephone," "radio," "solar panel," "cell phone," "thermometer," "thermostat"). Theory-laden though they are, we need those terms to describe facts of everyday, pre-theoretic experience. We therefore need them to state the pre-theoretic data that our theories are to model.

"But," it will be said, "anything that can be described *only* by taking for granted some theoretical belief is *ipso facto* not a *datum.* A truth? Maybe. A datum? No."

That may well be true. But, if so, our belief-system is so overrun with theoretical beliefs that no belief system free of such beliefs would bear even the slightest resemblance to it. Also, it is incoherent to suppose that there could exist a belief-*system* that wasn't replete with theoretical content. The difference between a mere belief, on the one hand, and a belief-system, on the other, is that belief-systems are intended to be explanatory, whereas mere beliefs needn't have this property. I believe that it's raining because I see that it's raining—*not* because, in an effort to model various observational data, I at some point *hypothesized* that it was raining. But psychoanalytic theory constitutes a belief-system, as opposed to a mere belief, because it is intended to be explanatory. This suggests (though it doesn't establish) that "belief-system" and "theory" are veritable synonyms. (Why the parenthetical hedge? "Because there are belief-systems that aren't of a theoretical nature. Religious belief-systems aren't theories." But even this is arguable.[236] Some would say, and I agree with them, that religious systems are proto-theories or para-theories—structures that make heavy allowances to emotional factors and make insufficient allowances to considerations of logic and evidence, but are, within these parameters, as theory-like as they could possibly be.) If, indeed, belief-systems are *ipso facto* theories, it makes no sense to speak of what our "belief-system" would be like if it were stripped of theoretical content. Long story short: although it's obvious that, on its own, (*) is perfectly coherent, it's far from obvious that (*) is compatible with our belief system as a whole or with any belief system that isn't radically different from it. This was the very point we set out to establish. It goes without saying that what is true of (*) holds *mutatis mutandis* of each of many other "deviant" hypotheses.

But even if, for some reason, this last point is denied, there is no way to deny our main point, to wit: even if (*) is logically consistent with our belief system, or at least with some not-too-mutilated version of it, the explanatory benefits of countenancing (*) in favor of the conventional explanation are grossly outweighed by the explanatory costs.

4.3 Synthesizing these points

It will help if, before moving forward, we take a moment to synthesize the points thus far made. There's no difference between committing the Gambler's fallacy and performing an induction by enumeration. At the same time, knowledge of invariable concomitances is obviously an extremely valuable source of knowledge. But it isn't, or in any case shouldn't be, through enumerative induction that one transitions from (1) "all phi's that I've known of were psi's" to (2) "the next phi I come across will be a psi." (In this context, it is to be assumed that that "x is a phi" doesn't analytically entail "x is a psi.")

The *actual* bridge between (1) and (2), supposing that there is one, lies in the fact that, given (1), it is reasonable to believe there to be some mechanism or natural law that requires (or at least disposes) phi's to be psi's. By itself, (1*) "all metal known to me has expanded when heated" provides no good reason to accept (2*) "all metal expands when heated." (1*) provides a good reason for accepting (2) *only* to the extent that (1*) provides a good reason for accepting (3*) "there is a lawful connection between a thing's being made of metal, on the one hand, and it's expanding when being heated, on the other."

4.3.1 Synthesizing these points (continued)

Before we continue, it's crucial to note that none of this begs any questions against Hume. Hume is saying: "given only that every metal object that Smith *has* known of has expanded when heated, it doesn't follow that metal that Smith has *yet* to encounter will (or is even likely to) expand when heated." We agree. Smith is guilty of the gambler's fallacy so far as he thinks otherwise. All we're saying is this: Smith's experience thus far *suggests* that "x's being heated metal provides some kind of nomic basis for x's expanding" *has* held for objects falling within the scope of Smith's experience.

We're *not* making the claim Smith's experience *entails* that some such connection has held. Such a claim would simply be false. And we're *not* making the claim that Smith's past experience gives him a good reason to think that metal objects not (yet) known to him will be more likely to expand or even that they'll be more likely than not to expand. Such a claim would beg the question against Hume. Finally, we're *not* making the claim that Smith's past experience gives him a good reason to think that there *continues* to exist any nomic connection between a thing's being heated metal and its expanding. Echoing what we just said, such a claim would beg the question against Hume: for Hume's very point is that a thing's happening in the past is no guarantee of its continuing to happening the future. And we cannot, without begging the question against Hume, assert that a law's having held in the past is evidence of its holding in the future.

So what *are* we saying? Only that Smith's experience gives him a reason to believe that such a connection *has* held, at least as far as metal objects falling within the scope of experience are concerned. And in making this obviously plausible claim, no questions have been begged against Hume.

4.3.2 Synthesizing these points (continued)

I propose that, armed only with this non-question-begging point along with the tautology that it is the purpose of explanations to eliminate unexplained explainers, we can show that Smith has reason to expect *future* metal objects to expand when heated.

We've agreed that Smith's experience does give him some reason to think that, at least as far as objects thus far known to him are concerned, a thing's being metal provides some kind of causal or nomic basis for its expanding when heated. Let P be this causal principle. P is of doubly limited scope. It applies only to *past* objects *and* it applies only to past objects that Smith knows of. Let's suppose that Smith accepts P.

Smith has two options. (I am not counting a refusal draw any inferences as an option.) (i) Hold that provides no reason to believe that, outside Smith's past, personal experience, a thing's being a metal provides a basis of some kind for its expanding when heated. (ii) Hold that P does warrant some such generalization. One such generalization is: "that connection holds for all past objects, whether Smith knows of them or not; but it doesn't hold of future

objects." Another is: "that connection holds for objects that Smith *will* know of, but not for those that he won't know of." But, of course, the most natural, not to say the most rational, generalization is this:

(P*) "That connection holds *generally.* It holds no less for objects not known by Smith than for those known by him, no less for future objects than for past."

Let's suppose that Smith chooses *not* to accept P* or any other generalization of P. How, in that case, would the world have to be for Smith's belief system to hold? Metal objects that Smith became aware of would suddenly undergo a profound change: they would go from being things that, when heated, didn't expand to being ones that did. And for this change to occur, each atom of every metal object known to Smith would have to change in fundamental ways. Given an acceptance of P, it may seem conservative not to accept any generalization of P. But it's the very opposite. If you accept but reject any generalization thereof, you're saying that objects miraculously undergo profound microstructural changes whenever you know of them and undergo equally profound changes when you stop being aware of them. And *that* hypothesis is not a conservative one.

And, quite clearly, it's a bad one. Why is it bad? Because it creates more unexplained explainers than it eliminates—because it creates more questions than it answers.

4.3.3 Synthesizing these points (continued)

Of course, *not* accepting a hypothesis isn't the same as rejecting it. *Not* accepting some generalization of P isn't the same as rejecting it. And it's obviously a good idea in many cases to hedge—to wait for evidence. But that's irrelevant. The question is: given that there is enough information to warrant an acceptance of P, is there enough information to warrant acceptance of any generalization of P? If we answer by saying "no," then we're saying that, given only the evidence at our disposal, it's no less likely than not that objects systematically undergo massive changes at the atomic level whenever they come or cease to be known by Smith. And *that* position, as we saw, creates more mysteries than it eliminates. So despite first appearances, there's nothing conservative about refusing to countenance any

generalization of P. And such refusal embodies acceptance of a principle that is straightforwardly, because tautologously, false, namely: “there’s no reason to prefer explanations that explain a lot and create little that has to be explained to those that explain a little and create a lot to be explained.”

4.3.4 Synthesizing these points (continued)

These points are easily generalized. It’s no less absurd to suppose that, in the last nanosecond, all metal objects have undergone profound microstructural changes than it is to suppose that all metal objects known by Smith are microstructurally very different from those not known by him. This point provides the non-circular justification for UP that Hume sought but couldn’t find.

Let’s suppose that Smith has reason to believe (i) that, where metal objects that he *has* known of are concerned, a thing’s being heated provides some kind of basis for its expanding. Let’s also suppose that Smith holds (ii) that this principle *won’t* hold for metal objects that he *will* encounter (i.e., for metal objects that he hasn’t encountered but will and for those that he has already encountered and will encounter again). In holding (ii), he is *ipso facto* supposing that the laws of nature will abruptly change. And in positing this one change, he is positing innumerably many others, as we will now see.

A metal object’s being heated involves innumerable microstructural changes, and so does its expanding. The macroscopic supervenes on the microscopic. Given two objects that are microstructurally alike—that consist of the same particles interrelated in the same way—there is no possibility, even a purely logical one, that those two objects should differ at the macroscopic level. By the same token, given two objects that differ at the macroscopic level, there is no possibility, even a purely logical one, that those two objects should *fail* to be microstructurally different. A metal object’s being heated thus consists in its undergoing many microstructural changes, and so does its expanding. And if a metal object’s being heated causes it to expand, it is in virtue of the fact that the first set of microstructural disturbances result in, or are themselves constitutive of, the second set.

So given a law—even one that was confined to objects that Smith has encountered—to the effect that metal expands when heated, in order for that

law to be annulled, each of innumerably many microparticles would suddenly have to undergo quite dramatic changes. So in assuming that, all of a sudden, a law that once held ceases to do so—or, what is the same thing, in assuming that the future, starting now, is governed by different laws from the past—one is positing not *one* change, but innumerably many. For, in making that assumption, one is (implicitly) supposing that each of the innumerably many microparticles will, all of a sudden, stop being governed by one set of laws and will start being governed by another. And one is thus creating what, for practical purposes, might as well be infinitely many unexplained explainers.

But not a single one of these unexplained explainers is created by the hypothesis that metal objects that we will encounter are governed by the same laws as those we have encountered. So given the truism that one explanation is better than another if, other things being equal, the first creates fewer unexplained explainers than the second, it follows that the rational move, at least until further evidence comes in, is to suppose that those microparticles will continue to be governed by the same laws that used to govern them.

According to what is now the conventional wisdom, which coincides with the viewpoint that Hume bequeathed to us, the conservative move is the agnostic one: stick to the evidence; generalize as little as possible; don't assume that what you don't know resembles what you do know. But the so-called conservative view is really *less* conservative than the alternative. (The alternative is: "assume that what you don't know *does* resemble what you do know.") The so-called conservative view posits giant schisms in the natural order and, therewith, masses of unexplained explainers. The alternative view does not.

5.0 Closing the argument

We can now succinctly say why Hume's argument fails.

Step 1: It's an *analytic* truth that *ceteris paribus* the better explanation is the one that creates fewer unexplained explainers. Let's refer to this principle as "MC" (short for "minimization of causal anomalies"). Whereas UP, if it can be known at all, can be known only on inductive grounds, MC is analytic and, unlike UP, is therefore not to be known on empirical or, consequently, inductive grounds. Thus, MC, unlike UP, can be noncircularly assumed in the

context of a defense of inductive inference.

Step 2: It's a datum that various concomitances are to be found *among the objects known to one.* It's a datum that, so far as the object's known to (say) *Smith* are concerned, putting a hand in boiling water causes intense pain, whereas putting that same hand in cool water causes intense relief.

Step 3: One creates more unexplained explainers by assuming that the unknown differs from the known than by assuming that it does not.

Step 4: If one remains agnostic—if one refuses to say, one way or the other, whether the unknown resembles the known—one is committed to the incoherent position that there's no reason to prefer a hypothesis that creates few unexplained explainers to an otherwise comparable one that creates many. (The reason this view is incoherent is that for one explanation to be preferable to another just *is* for it to do more in the way of explaining, and thus less in the way of creating unexplained explainers, than the other.)

Step 5: Given MC, and given Steps 1–4, it follows that there is an analytic and therefore *non*-inductive reason to hold that, other things being equal, the unknown resembles the known.

Conclusion: Hume is wrong; Hume's argument fails. This is a direct consequence of Step 5. Given the legitimacy of MC, it's obvious why, though it may initially appear cautious and conservative, skepticism about the external world is actually an extravagant and irrational position. One's sensory experience is replete with discontinuities. Every time you open your eyes or turn your head, you are flooded with sensory experiences that are totally unlike those that you were experiencing a moment earlier. According to the non-skeptic, these new experiences don't come into existence *ex nihilo;* they are merely links in chains of events whose other links you don't perceive. The skeptic says that it's no more rational to believe this about our sensory experiences than it is to believe that those experiences come into existence *ex nihilo.* In saying this, the skeptic is saying, in effect, that it's just as rational to accept a theory that creates more anomalies than it gets rid of than it is to accept an otherwise comparable theory that gets rid of more anomalies than it creates. Since it's an analytic truth that, given two otherwise comparable theories, the one that leaves us with fewer anomalies is the one that ought to be preferred, it follows that the skeptic's position is analytically

false. Of course, it isn't analytic that our sensory experiences are correct. But it *is* analytic that it's rational to suppose our sensory experiences veridical *to the extent that doing so* leaves us with fewer causal anomalies than we're otherwise left with. The skeptic's denial of this embodies an analytically false conception of explanation.

Chapter 13

Empiricism and Its Limits

1.0 Empiricism vs. rationalism

We obviously acquire a great deal of knowledge through "sense-perception" (i.e., through sight, hearing, touch, and so forth). According to a doctrine known as "empiricism," *all* knowledge is derived from sense-perception.

According to a view known as "rationalism," some knowledge is acquired entirely through the use of one's ability to *reason.*

Rationalists almost never hold that *no* knowledge is acquired through sense-perception. They hold only that reason, as opposed to sense-perception, is the vehicle through which *some* knowledge is acquired.

Rationalists typically hold that knowledge acquired in this way is very important—it isn't trivial.

Some hold the view that there is knowledge that is acquired neither through the senses nor through reason. I don't wish to dismiss this view. Maybe it's correct. But there is an apparent problem with it. Any case of knowledge is a case of *justified* true belief. Given a belief that isn't acquired through the senses *or* through reasoning, the question arises: what could possibly justify it? And there's no obvious answer.

In any case, in this chapter, it will be assumed, if only to simplify exposition, that the only viable sources of knowledge are reason and the senses and, therefore, that the only viable options for epistemologists are rationalism and empiricism. But I wish to stress that I am genuinely open on the question of whether there is non-perceptual, non-rational knowledge. (I'm

inclined to think that there is.)

Those who believe that there exist non-spatiotemporal entities are necessarily rationalists. Give or take a few nuances, empiricism is the view that you don't know it if you don't see it. We see property-instances, but not properties themselves. Therefore, empiricists either deny the existence of properties themselves or they hold that properties are identical spatiotemporal entities. Those who take the latter view are forced to view properties as being composed of their own instances; so the property of being wet is identical with some object composed of all wet things, or some such. Therefore, empiricists don't believe in *uninstantiated* properties. Therefore, anyone who does believe in such properties, or who believes that properties are not spatiotemporal entities, is a rationalist.

Those who believe that properties are non-spatiotemporal are *Platonists.* Therefore Platonists are rationalists. A Platonic metaphysics requires a rationalist epistemology. Plato himself accepted a very extreme form of rationalism.

Being one that has deep roots in common sense, empiricism is a very old doctrine. But empiricism was first rigorously developed by John Locke (1632–1704), George Berkeley (1685–1753), and David Hume (1711–1776). The doctrines put forth by these authors form the basis of modern philosophy. Hume's beliefs about *causality* and *inductive inference* are outgrowths of his empiricism. These views are thoroughly examined in Chapters 12 and 17.

Berkeley's belief that objects are identical with our perceptions of them is an outgrowth of his empiricism. Berkeley's ingenious arguments for this outrageous doctrine are examined later in this chapter.

John Locke's plausible position that universals are "the workmanship of the understanding"—that *we* create them, so as to make sense of the world—is a derivative of his empiricism. That doctrine is examined in the present chapter. (It is also examined in Chapter 2.)

The first great rationalist was Plato (428 B.C.–348 B.C.). (Plato's views are defended in Chapter 2.) Other rationalists are: Augustine (354–430), Descartes (1596–1650), Leibniz (1646–1716), and Gottlob Frege (1848–1925).

The first great empiricist was Aristotle (384 B.C.–322 B.C.), whose views,

along with Plato's, are also considered in Chapter 2. Other empiricists are J.S. Mill (1806–1873), Rudolph Carnap (1891–1970), and W.V.O. Quine (1908–2000).

In this chapter, we will find that empiricism is not in its strictest form a defensible doctrine and that, consequently, at least some knowledge has a non-perceptual basis.

There are, as we will see, two important corollaries. First, many commonsense-based views about the internal (psychological) and external (physical) worlds must be jettisoned. Second, sober scientific methodology actually demands an acceptance of principles that would strike common sense as decidedly unscientific. We will see that many concepts that cannot be adequately understood along strictly empiricist lines can nonetheless be correctly understood only *interms of* what empiricism has to say about them.

2.0 The particularist nature of sense-perception and the need for universals

Sense-perception discloses only *particulars* to us—particular objects having particular property-instances. But we understand and describe the world in terms of categories or, as they are sometimes called, "universals." These can be thought of as natures, essences, or "ways of being," that, although *exemplified* by the particulars that we encounter in sense-perception, are not themselves to be sense-perceived, and, though they presumably exist, do not exist in space-time.

Because they are non-spatiotemporal, those of an empiricist mindset regard universals with extreme suspicion. In fact, empiricism is, nearly enough, the denial that such things exist. Why? The empiricists say: what we can legitimately believe is what we can sense-perceive or know on the basis of extremely conservative inferences on the basis of what we can sense-perceive.

The problem is that there is no way to understand the world except in terms of universals. To understand things *is* to know how, in virtue of having such and such characteristics (i.e., in virtue of falling under such and such universals), they are subject to thus and such principles or laws. "Why did Max steal the money? Because Max is greedy and amoral and he believed it

to be to his strategic advantage to steal the money." Max's behavior is explained by showing how, given the operative conditions, his having those characteristics makes it inevitable that he will act in that way. This explanation presupposes the existence of many different universals. (Among them are those denoted by the terms "greed," "amorality," "money," and "theft.") This explanation is not unique in this respect: *all* explanations presuppose the existence of universals. The very idea of an explanation that didn't do so is an incoherent one.

Universals are outside of space-time. For this reason, they cannot have effects on anything. (Causal relations hold only among spatiotemporal entities.) One cannot sense-perceive an object without being affected by it. (It is only because your ocular surfaces are disturbed by light-rays bouncing off of the page in front of you that you can see it.) It follows that universals cannot be sense-perceived. It also follows that no *evidence* of them can be sense-perceived. This is because x is evidence of y only if x and y are *causally interconnected* in such a way that, given x's existence, y's may reasonably be assumed. (Smoke is evidence of fire because smoke and fire are causally connected in such a way that, given the presence of smoke, the presence of fire may be inferred.) It is therefore impossible as a matter of logic to find any *observational* basis for the existence of universals. This means that, if empiricism is correct, we cannot possibly have any good reason to posit the existence of universals. Given how endemic universals are to thought, language, and explanation, the right move is to reject empiricism.

2.1 The empiricist response to these problems

Empiricists have responded by saying that, although *instances* of universals obviously exist, universals *per se* do not. (Remember that, in this context, "universal" is synonymous with "property.")

The problem with this view is that, as previously noted, every sentence and every thought demands the existence of at least one universal.

Also, some basic logic shows that universals exist. Consider the following inference: "Bob and Ted are both human; therefore, there is some characteristic that they have in common." That is a valid inference; and the conclusion affirms the existence of a universal. ("Characteristic" is

synonymous with “universal.”)

Finally, many statements straightforwardly presuppose or affirm the existence of universals; and were it not for statements of this kind, many obvious truths couldn’t be expressed. Consider the following statements:

(i) Hitler and Stalin had a lot in common;

(ii) Smith has what it takes to be a great pianist, but he doesn’t have what it takes to be a great composer;

(iii) Even though Smith is different from Jones, that has nothing to do with the fact that they don’t like each other.

The meanings of these sentences are, respectively:

(i*) There are many properties such that, given any one of them, Hitler had it and so did Stalin;

(ii*) There are properties such that, if one has them, one is a great pianist and such that, moreover, Smith has them; and there are properties such that, if one lacks them, one isn’t a great composer and such that, moreover, Smith lacks them;

(iii*) There are properties such that Smith has them and Jones does not, but this fact isn’t responsible
for the fact that they don’t like each other.

(i)–(iii) cannot be translated into statements that don’t presuppose or affirm the existence of properties. So it isn’t really an option to hold that properties don’t exist.

As soon as one grants that anything resembles anything, one must grant that universals exist independently of thought. For two objects to be similar in some respect is for there to exist some property that they both have. For two objects to have the same shape is for there to exist some morphological property such that each object is an instance of that property. It’s obviously possible for two objects to have the same shape in a world devoid of sentient beings. Therefore, there exist properties in worlds where there are no sentient beings. Therefore, properties aren’t created by such beings and aren’t otherwise dependent on them.

Properties are identical with *categories.* For x to fall into a given

category is for x to have a given property. (For x to fall into the category of round things is for x to have the property of being round.) Given that properties exist independently of thought, the same is true of categories. Thus, contrary to what some empiricists (e.g., Locke) say, categories are not devices that human beings created in order to expedite thought.

It is easy to find independent corroboration for this view. Let each of x and y be a pint of pure, clean glacial water, and let z be a pint of gasoline. The chef or chemist or auto mechanic who chooses to co-categorize x and z, as opposed to x and y, is in for some very unpleasant surprises. Why? Because his categories don't cut "nature at the joints," to use Francis Bacon's expression. In other words, there is some relevant property that x and y have in common that x and z *don't* have in common, and this person's system of categories embodies an ignorance of that fact.

3.0 Non-perceptual knowledge a prerequisite for perceptual knowledge

It's not as though tables, rocks, trees, etc., create little images of themselves that they deposit in our hitherto empty crania. Our minds have to supply the "paint," so to speak, with which our mental representations of those objects are painted.

In fact, our minds must supply *everything.* External objects don't *supply* our minds with anything. All they do is tell our minds when to deploy what is already in them. External objects don't stock our minds; they *activate* what's already in them. A rock will never experience an image. It doesn't matter how many light beams you fire at it. There's nothing in it to be brought forth; there are, so to speak, no images in there to begin with. For similar reasons, the acoustical disturbances that, in a human, would lead to knowledge of English or of philosophical principles will have no comparable effect on a rock. Thoughts are taught by being elicited, not by being deposited. This is the most fundamental problem with empiricism.

A precondition for sense-perceiving anything is that we be able to represent it as having a location (relative to us) in space and time, that we be able to represent it as having a certain shape, state of motion, color, etc. Although we learn a great deal through perception, what we *don't* learn

through it are the concepts—space, time, motion, shape, color, etc.—out of which sense-perceptions are constructed. So those things, and all that having them involves, must be in the mind *prior* to sense-perception, as a necessary *pre*condition for it.

These considerations expose what I believe to be the deepest confusion of which empiricism is guilty. Empiricists fail to distinguish between two very different roles that sense-perceptions can have: they can have a *triggering* role and they can also have an *information-transmitting* role.

3.1 Sense-perception as transmitter of information about the external world

You see a dog. On that basis you have the belief that there is a nearby dog. Here your eyes transmitted information to you about the external world. They "told" you, so to speak, that there was a dog in a certain place. Your could not possibly have learned this fact through pure logic-chopping or conceptual analysis. You could learn it only through the senses (either by seeing the dog or by sense-perceiving some other sort of evidence of the dog). In this context, then, information about the external world is being transmitted to you through your senses. Your senses (specifically, your sense of sight) is thus functioning in what we might call an "information transmitting" capacity.

3.2 Sense-perception as trigger

Sense-perceptions have a triggering role when they provide the recepient with new information *not* by telling him something about the external world, but by activating or "triggering" some already existing cognitive faculty.

You are in a geometry class. The teacher draws something that is supposed to represent a triangle but only vaguely looks like one. He draws various lines through it, around it, etc. And from these diagrams you learn various facts about triangles.

What did you actually *see?* What did your eyes show you? A triangle? No—only a few uneven lines that formed something that vaguely looked like a triangle. How did your seeing this non-triangle get you to know that, for example, the area of a triangle is 1/2 base times height? Since you didn't see

any triangles, you didn't see any triangles' height, base, or area. The things you saw were symbols for things that had to be supplied to you by something other than your senses—that had to be supplied to you by your intellect.

This is not to say that the diagram was useless. Plainly it was not. It may be that, but for seeing it, you wouldn't have learned anything about triangles. But your seeing the diagram led to your knowing those truths about triangles *not* because any triangle, or any fact about triangles, was any part of the information imparted to you by your perception, but rather by *triggering* or *activating* certain conceptual faculties of yours. In this context, your senses are (to the extent that you are learning actual geometrical truths as opposed to mathematically irrelevant facts about the blackboard, chalk-marks, etc.) providing you with knowledge by getting you to acquire knowledge in some *non*-sensory way, viz. by getting you to grasp, and analyze relations among, concepts.

The proof that your visual perception of the diagram had only a *triggering* role, and that it did not itself contain any information relating to what you learned, is that, if you were to forget ever having seen that diagram —were to forget what it looked like, who drew it, where you saw, or even *that* you saw it—it would have no effect on the geometrical truths that you learned that day. The reason is that those geometrical truths (e.g., area of a triangles = bh/2) were simply *not* part of the content of what your senses were telling you.

By contrast, if you forget ever seeing the dog (described in the last section), you *will* forget the information that, on that occasion, your senses led you to have. The reason is that, in that context, your senses weren't serving the purpose of jogging your ability to acquire knowledge in some *non*-sensory way (viz. through rational insight) but were themselves the bearers of the information learned.

3.3 An anti-empiricist consequence of the distinction between triggering and information transmitting

According to strict empiricists, *everything* that is learned, including purely

logical principles, is learned through sense-perception. When this appears to happen, it's because sense-perception is simply "turning on" some non-sensory mode of knowledge acquisition. "Truths of reason," as Hume called them, aren't learned through sense-perception, even though sense-perception may awaken the non-sensory faculty through which we do become aware of them.

Some empiricists are aware of this objection, and they generally deal with it by saying one of two things:

(1) Truths of reason *really* are truths about the irregular figures on blackboards, that truths of arithmetic *realyl are* truths about what happens when apples are put on top of other apples, etc.

(2) The so-called "truths of reason" are just linguistic or typographical conventions. "1+1=2" is an empty, purely definitional truth. "2" is *defined* as "1 + 1," and "3" is *defined* as "1 + 2." Every other fact about mathematics can be reduced to definitions, linguistic conventions; and all of mathematics turns out to be typographical. (As a matter of typography, you can replace "1 + 1" with "2," just as you can replace "five" with "5.") And all the so-called "truths of reason" that lie outside of mathematics are thought to be explained in the same way.

3.3.1 Why (1) is wrong

Your knowledge that 113 + 876 = 989 isn't empirical; you didn't see 113 pears being counted, then see 876 pineapples being counted, and then see the whole lot be counted. (And even if you did have such an experience in connection with those particular numbers, there are many other pairs of numbers whose sums—not to mention quotients, products, differences, etc.—you know, or could readily figure out, but in connection with which you've never seen anything like the situation just described.) Your knowledge of arithmetical truths is based on a grasp of non-empirical principles. Even if (what I think is necessarily false) you learned that 1 + 1 = 2 from watching one thing be put on top of another, you still would not, on that basis alone, have any way of generalizing that knowledge—of inferring the other mathematical truths you know. To make the needed inferences, you'd need

access to rules of inference that you couldn't possibly know about if your knowledge were strictly sense-based.

Another reason why (1) is wrong: Even one's knowledge that 1 + 1 = 2 isn't known empirically. As David Hume made clear, observation tells you how things are, not how they should be. It reveals facts, not norms. But one's knowledge that 1 + 1 = 2 has normative force. If two people enter what you thought was an empty house, and three people leave, you don't reconsider your belief that 1 + 1 = 2. You take that belief as fixed and, in light of it, reconsider your belief that the house was in fact empty. If it turned out that, according to a certain scale, rocks R and R* had a collective weight of more than 2 pounds, even though you previously thought that each weight no more than 1 pound, you'd take your belief that 1 + 1 = 2 as fixed, and would revise some other belief of yours (e.g., your views concerning the rocks' individual weights or the accuracy of the scale or possibly even the laws of physics).[237] We interpret observations in light of the principle that 1 + 1 = 2, not *vice versa.*

3.3.2 Why (2) is wrong

Conventions are indeed arbitrary. But whether the conventions one adopts are *consistent* with one another is not itself a matter of convention. It's a matter of convention that "1" denotes 1, that "+" denotes addition, and it may even be strictly a matter of convention (though I think not: see below) that "2" refers to the same number as "1 + 1." But once these conventions have been laid down—once convention has fixed what "1," "+," "=," "3," etc. mean—various facts that are *not* among those conventions follow. For example, *given* our convention of referring to "1" as 1, to "+," as plus, etc. it follows that there is no pair of whole numbers p and q such that $(p/q)^2$[238] = 2 (in other words, 2 has no rational square root). But it's not a matter of *convention* that "2 has no rational square root" is true. It is a non-conventional consequence of our conventions. (Nor—to make a different point—is it a matter of convention that what is *meant* by that sentence is true. The *fact* that 2 has no rational square root isn't a consequence of our conventions; it has nothing to do with conventions, as we'll discuss in a moment.) So even though our conventions don't directly say anything about the permissibility

of the following sentence—"there *is* a pair of whole numbers p/q such that $(p/q)^2$"—those conventions *prohibit us* from adding that sentence to those same conventions. Were we to do so, our conventions would be *inconsistent.* In general, not all truths of reason (or any, actually, as well now see) can be conventionalized, since consistency among conventions is not itself a matter of convention.

3.3.3 Another reason why (2) is wrong

Also, we have to distinguish sentences from the things they express. For argument's sake, suppose that "2" is *defined* as "1 + 1," and suppose that "=" is defined in the normal way. In that case, the *sentence* "1 + 1 = 2" will indeed be a matter of definition. But that leaves it completely open *which* truth that sentence expresses. For suppose that, as a matter of convention, "1" referred to 8, and "1 + 1" referred to "7 + 1." Or suppose that as a matter of convention, "1 + 1" referred to Superman and "2" also referred to Superman. In each of those cases, the *sentence* "1 + 1 = 2" would be true by convention, but each would express a truth very different from the one actually expressed by "1 + 1 = 2." The truth *actually* expressed by that sentence concerns the number one—not Superman, not the number eight—and it says of it that, when added to itself, the result is two—not Superman, not eight.

Could "1 + 1 = 2" express *that very truth*—the one just described—and be false? No. If a sentence affirms that one and one make two, then—because of facts about one and two and addition and equality—it's going to be true. So what makes "1 + 1 = 2" true isn't convention, and is instead non-conventional facts about numbers. To be sure, conventions are involved: it is a matter of convention that "1" refers to one, "+" refers to addition, etc. And it follows from those conventions that "1 + 1 = 2" says that one, added to itself, yields two. But it's a matter of mathematical, non-conventional fact that one plus one equals two.

4.0 Why pre-perceptual and, therefore, non-empirical knowledge is a prerequisite for empirical knowledge

Innate cognitive structures are needed to have sense-perceptions and also to learn from them. Having a sense-perception means having a certain kind of mental representation of an object. Not all mental representations of objects are sense-perceptions. Mental images aren't, even though they resemble them. Concepts aren't, and they don't resemble them. It's a delicate matter exactly what kinds of representations are perceptual: a necessary condition, certainly, is that they be caused by their objects—your perception of Smith must be caused by Smith (by light bouncing off of him). It also seems necessary that (with certain qualifications) your perception "co-vary," or be capable of co-varying, in real time with the thing perceived. If you're looking at Smith, and he moves, you see him move (provided certain conditions are met; e.g., the movement isn't infinitesimally small). It's unclear whether there are other conditions that have to be met. (There probably are.)

In any case, perceptions are mental representations. And they represent their objects in four dimensions, and as having various chromatic (color-related), kinematic (movement-related), dynamic (force-related, causal) properties. Unless we take the implausible view that no intelligence is involved in the processes mediating between the somatic disturbances that lead to sense-perceptions, on the one hand, and those sense-perceptions themselves, on the other, none of this would be possible unless your mind, at some level, had some kind of pre-perceptual understanding of what it is for two things to be spatially, dynamically, etc., interrelated. So, in all likelihood, sense-perception *presupposes* knowledge.

4.1 Why non-empirical knowledge is a prerequisite for post-perceptual, inferential knowledge

We couldn't do very much with the information that sense-perception gave us unless we had non-empirical knowledge. First of all, your senses operate in the present. You are not seeing, touching, etc., whatever you're not *now* seeing, touching, etc. At most, you are remembering it. In order to apply what you remember to a current situation, you must see how what you remember *bears* on your current situation. Let's take an example that, because it concerns such a pedestrian inference, shows how great the scope of this point is. You remember seeing somebody turn the key in their car and then driving away. You are now in front of a car with a key in it, and you want to drive it.

Obviously you can apply what you remember to this situation. Let us suppose that you do so. So you think: "on *that* occasion, turning the key started the car; so the same thing will happen on *this* occasion. So, given that I want to start the car, I'll turn the key." Sense-perception (plus memory) *has* shown you that *on the first occasion* a certain car started after the key in it was turned. And, supposing that you turn the key in this car (the one in front of you), sense-perception will show you that another car started after the key in *it* was turned. But sense-perception did not show you, and could not possibly have shown you, that there was a *relationship* between these two situations; it could not show you that, *because* the car in the first situation started after the key was turned, the same would be true of the car in the second. Your senses apprise you of situations; they thus give you *intra* situational information. They don't give you *inter* situational information. You can't see or otherwise sense-perceive relations between situations—between an event you saw a week ago and one you're seeing right now. Thus, any knowledge that you have intersituational relations is partly (though obviously not entirely) non-empirical. Therefore any knowledge of *dependence*-relations holding between distinct situations is, at least in part, not strictly sense-based. Since it is only by virtue of knowing of such relations that one can know anything other than what one's senses are currently telling one, one can know extremely little if all knowledge is sense-based, and what one can know, one can know only for a fleeting instant.

4.1.2 Why non-empirical knowledge is a prerequisite for post-perceptual, inferential knowledge (continued)

Hume said that there is *no* legitimate inference from "that car (the one I saw two days ago) started when the key was turned" to "this car (the one I'm looking at now) will start if the key is turned" and, in general, that there is no legitimate inference from *any* proposition concerning the past to *any* proposition concerning the present or future. Hume acknowledges, of course, that past experience *makes* people form certain beliefs.[2] But his point is precisely that those beliefs, lacking as they do any legitimate basis, are not knowledge; they're just compulsions (albeit ones that, as he grants, have thus

far served us well for the most part).

Although I believe that knowledge of the past *does* warrant beliefs about the present and future, there is an important truth in what Hume is saying. Hume has made it clear that *f i* one permits oneself to use *only* such knowledge as one has acquired through the use of one's senses, then one can't infer anything about anything from anything. Inferences are about bearing relations. (More precisely, if one legitimately infers Q from P, then (i) P supports Q, (ii) one knows that P supports Q, and (iii) one's acceptance of Q, given P, is driven by one's knowledge that P supports Q.) If something about situation X is inferred from something about situation Y, that inference, if legitimate, must accord with some principle that is concerned with relations holding *between* distinct situations. And such relations, obviously, cannot, at least not as a rule, be sense-perceived. To sum up, all knowledge that isn't strictly sense-based—and this means all knowledge that involves any element of inference and, therefore, all theoretical knowledge (all theoretical knowledge is inferential, but not *vice versa*)—one must have knowledge of intersituational dependencies; and, since they can't be sense-perceived, one's knowledge of these dependences cannot be strictly sense-based.

5.0 The biggest problem with empiricism: not all thoughts are pictorial

What does perception give you? Images. (Not just visual images, of course, but also acoustical images, tactile images, etc.) If empiricism is right, this means that any thought that is not itself an image, or that contains information that cannot be encoded into an image, is unwarranted. But most thoughts, if not all of them, contain just such information. In other words, most thoughts, if not all of them, contain *non-iconic* information. For example, there is no image such that, in virtue of experiencing that image, one is thinking:

(S) if Smith had fallen off the fence, he would have broken his leg.

What sort of image would represent S? Obviously you could have an image of Smith falling from a fence, and you can have an image of Smith breaking his leg. But S is to the effect that Smith's falling would *involve* his breaking

his leg; it affirms the existence of a certain *dependence relationship.* And nothing could be an image of such a relationship. There can be images of rocks, trees, explosions, fallings, and waterfalls—but not of dependence relationships.

5.1 No iconic representation of dependence relations

Of course, you could create a system of symbolism *involving* images that represented that supposed truth (e.g., you could stipulate it would be representing by putting an image of Smith falling off the fence on top of an image of Smith with a broken leg). But that wouldn't be an *image* of its being the case that Smith would have broken his leg if he'd fallen off the fence. It would be an image of Smith falling off the fence on top of an image of Smith having a broken leg. Relative to some convention, the one image's being on top of the other would *represent* that dependence relation. But it wouldn't be an *image* of it. Relative to certain conventions, the expression "Socrates" represents Socrates, but it isn't an image of him. It's a non-iconic representation of him. And putting one image on top of another is no more a case of producing an image of a dependence relation than saying "Socrates" is producing an image of Socrates.

But even if there could be some image of this dependence relation, there would *still* be no image that had S for its content. There is a difference between believing S, on the one hand, and:

(S*) Smith broke his leg *because* he fell.

S* presupposes that Smith fell off the fence and that he broke his leg. S presupposes neither. In fact, it presupposes that Smith did not fall off the fence. So an image of S* wouldn't be an image of S, and an image of S would have to be an image to the effect that *interalia* Smith did *not* fall off the fence. But there can't be images of negative propositions. There is no image such that, by virtue of having it, you are thinking that Smith did *not* fall off the fence. You could produce an image with a big X on it of Smith falling off the fence. But that wouldn't be an image of Smith not falling off the fence. It would be an image with a big X on it of Smith falling off the

fence; or maybe it would be an image of Smith falling off a fence in the vicinity of a big X.

Disjunctive propositions are really conditional ones. *Either P or Q* is interchangeable with *if not P, then Q.* It follows that no disjunction can be imaged.

5.2 Iconic information too specific in some respects and also too unspecific in others

First a terminological point: By a "general" thought, I mean one that concerns some unspecified thing (e.g., *somebody or other cured cancer*) or that concerns all things in a certain category without considering any one of them in particular (e.g., *most people have at least some decency*).

Suppose you believe that the cat is on the mat. Any *image* of a cat on the mat will include information not included in that thought (e.g., it will include information concerning the cat's shape or color, or concerning the mat's shape or size).

By the same token, any thought will contain information not included in any image. An image of a cat on the mat is indistinguishable from an image of a perfect cat-impostor that is on a mat. So if the thought that *the cat is on the mat* were identical with an image, it would be identical with the thought that *a perfect cat impostor is on the mat.* But those propositions are distinct. [239]

Also, general thoughts couldn't possibly be identical with images. Suppose you think that something or other is on the mat. An image can't just be an image of something or other on the mat; it must be an image of some particular thing on the mat. So if the thought that something were on the mat were identical with an image, it would be identical with the thought that (say) Snoopy was on the mat. But those thoughts aren't identical.

Similar arguments show that other general thoughts can't be identical with images. Consider the thought that *all* birds have feathers. An image of *all* birds—an image that depicted every single bird in existence—would be indistinguishable from an image of (say) 84 billion birds. But the thought *84 billion birds have feathers* is different from the thought that all birds have

feathers. So neither thought is image. Obvious adaptations of this argument show that no image can be identical with the thought that most birds/many birds/all possible birds have feathers.

So, really, *no* thought that can be expressed by a sentence is identical with an image. This shows that many thoughts aren't images and, since sense-perceptions are images, therefore have contents very different from those of any sense-perception.

5.2.1 Berkeley's unwitting anti-empiricism

Ironically, it was that greatest of empiricists, George Berkeley, who made it clear that no general thought cannot be identical with images, the reason being that any image is specific in ways that any one of those thoughts won't be. (Berkeley didn't make the correlative point that thoughts are specific in ways that images are not.) Berkeley saw that *if* all thoughts are images, then there are certain thoughts we can't have. Berkeley also saw that, if empiricism is correct, thoughts must be images. Finally, Berkeley saw that, given these two points, empiricism is inconsistent with the presumption that we can have general thoughts. But instead of rejecting empiricism because it has this absurd consequence, Berkeley held that, indeed, we can't have general thoughts. This position, in addition to being obviously wrong, is self-defeating, since the thought that there are no general thoughts is itself a general thought.

Berkeley qualified his position by saying that, although we couldn't *directly* grasp general propositions, we could grasp non-general propositions that we *took to stand for general ones.* But one can't take x to mean y unless one grasps y. So one can't take a non-general proposition to stand for a general proposition unless one grasps the latter. (See Sections 8–21 of Berkeley's "Principles of Human Knowledge.")

5.3 Knowledge never wholly identical with mental imagery

As we observed, if all knowledge is strictly sense-based, then any case of knowledge is a case of some sequence of images running through one's mind.

But *no sequence of mental images is knowledge.* Such a sequence may serve as a partial *basis* for knowledge (e.g., if the image-sequence is a series of visual perceptions of a dog, then, partly on the basis of that sequence, you may know that there is a dog near you); and one can have knowledge *about* that sequence (e.g., you can know that ony a visually gifted person could experience such high-resolution mental images). Nevertheless, knowing something *always* involves grasping at least one thing in some way *other* than through an image and, therefore, in some way that, because sense-perception gives one nothing but images, empiricism cannot account for.

There's a difference between an image, on the one hand, and that image's being represented as *true,* on the other. (Philosophers use the word "veridical" to describe accurate perceptions, and they sometimes use the word "falsidical" to describe inaccurate ones. But I'll stick with "true" and "false.") Consider some painting (e.g., a painting of a woman picking flowers). There is nothing in the image itself that says: *this woman really exists, and she really is picking flowers.* Images are not self-referential; so they don't say "I am true" or "I am false." Thus, the image itself leaves it open to how it is to be taken. It could be taken as something true, something false, or something that is not meant to be taken as true *or* as false.

A corollary is that there is no image such that, merely by virtue of experiencing it, one is grasping the concept of truth. To know something, one must believe it to be true. Believing something to be true thus involves grasping the concept of truth and, given what we just said, therefore involves grasping something that cannot possibly be grasped by experiencing an image. Since the senses provide us only with images, it follows that the concept of truth is one that is not be grasped through the senses. And since a grasp of that concept is a prerequisite to knowing anything, it follows that, contrary to what empiricism says, not all knowledge is strictly sense-based. In fact, it follows that, although a great deal of knowledge is obviously *largely* sense-based, *no* knowledge is *strictly* sense-based.

6.0 George Berkeley and the collapse of empiricism into idealism

George Berkeley (1685–1753) argued that "to be is to be perceived" (*esse*

percipi est). He held, in other words, that the things we believe to be revealed to us *through* sense-perception (e.g., rocks, trees) *are* sense-perceptions—or, more exactly, are collections of sense-perceptions.[240] So Berkeley held that book that you see before you is identical with the perceptions you (and others) have of it.

The view that physical objects *are* sense-perceptions is known as "idealism." It is so named because it is the view that external objects are "ideas," not because it embodies a spirit of optimism.

Even though, if our intuitions are to be given any credence, idealism is an abomination, Berkeley had a very good reason for accepting it. Given an acceptance of empiricism, one *must* be an idealist *if* one is to avoid saying that we have no knowledge of the external world at all. Berkeley saw this; and, being an empiricist who wasn't also a skeptic, he duly accepted idealism.

According to empiricism, if you don't know it through sense-perception, you don't know it. But what do know through sense-perception? "I know that I'm reading a book right now." No you don't. You know that, *if what your senses are telling you is the truth,* you're reading a book.

Your senses cannot themselves give you any good reason to believe that they're truthful. Unless you already have some assurance that your senses are truthful, no attempt on their part to *tell* you this deserves any credence. So, supposing that you *do* know that you're reading a book right now, this knowledge, although *largely* based on sense-perception, isn't *entirely* so.

In general, even if our senses are generally telling us the truth, this fact cannot *itself* be known through sense-perception. This doesn't mean that it cannot be known. (I believe it can be known that our senses are telling us the truth. See Chapter 11.) But it means that, if we are to come to know that our senses are telling us the truth, it cannot be entirely on the basis of what our senses themselves tell us.

An analogy may help. Uncorroborated testimony is close to worthless. Smith is a suspect in a crime. He gives you an alibi. (You're the detective assigned to the case.) You want to know whether Smith's alibi is correct. If you ask Smith "are you telling me the truth?," he'll obviously say yes. What

you want is *independent corroboration* for Smith's testimony. No person can self-corroborate; and neither can anything else, including sense-perception. If we are to have any assurance that our senses are telling us the truth, that assurance must come from some non-sensory source. But empiricism says that nothing coming from any non-sensory source deserves any credit.

Given this fact, there are only two ways for empiricists, such as Berkeley, person to go. (i) Deny that we know anything about the external world. (ii) Redefine the concept of an external object in such a way as to validate empiricism.

(i) is self-explanatory. It's skepticism. Some philosophers accept skepticism.[241] But for Berkeley, skepticism is not alternative. He all but takes it for granted that it's false. (We'll say why in a moment.)

This puts Berkeley between a rock and a hard place. Berkeley refuses to stop being an empiricist. And he also refuses to accept skepticism.

This leaves him with only one option. He must show that the presumption that our senses give us knowledge of the external world can be accommodated *within* the limits set by the belief that all knowledge is strictly observational.

There is only one way to do this. One must reinterpret statements about the external objects as statements about one's own perceptions. In other words, one must take the view that the statement "that rock weighs five lbs" has the same content as some statement about one's own perceptions.

The position that statements about external objects can be translated into statements about perceptions is known as ***phenomenalism.*** Any attempt to translate a sentence about the external world into one about one's perceptions is known as a ***phenomenal reduction.***

Phenomenalism entails idealism. If every truth about rocks, trees, etc., is a truth about perceptions, then rocks, trees, etc., are nothing if not perceptions.

Phenomenalism is false. It simply isn't possible to translate statements about rocks (etc.) into statements about one's perceptions. We'll see why in Section 8.0.

But phenomenalism must be accepted by any true empiricist who isn't also a skeptic.

6.1 Berkeley's argument that to be is to be perceived[242]

Our senses give us knowledge of the external world (or so we will assume for argument's sake). There is no non-sensory source of knowledge (same qualification).

Our senses give us conflicting reports. I'm near the house; my eyes tell me that it's big. I'm far away from it; my eyes tell me that it's small. The first report conflicts with the second. I leave a sauna: the room feels cool. I then go outside (where it's frigid, since it's Northern Canada in the winter), and come back in: the same room feels hot (even though, as my housemates tell me, the room hasn't changed temperature).

In *some* cases where sensory-reports conflict with one another, we can throw one of them out. Somebody puts LSD in my drink. I hallucinate a nine-headed goblin. My goblin-(mis)perception can be thrown out; I know that it's wrong. Why? Because the message it gives me doesn't fit in with the other messages my senses give me. It is on this basis that we throw out dream-images on this basis. Dream-images don't fit into the narrative.

But I cannot throw out either of my house-perceptions. Neither is in the same category as my goblin-perception. Only a tiny fraction of the discrepancies within the reports given to us by our senses involve a (mis)perception that is in the same category as my goblin-perception. In most case, all of the mutually discrepant perceptions must be accepted. When, while I'm in an airplane, my eyes tell me that people on the ground are the size of ants, I cannot reject the testimony of senses. (I'm not dreaming or hallucinating.) Nor can I reject it when my eyes tell me that they're the same size as myself. During what follows, when we talk about "mutually discrepant perceptions," we *won't* be referring to discrepancies involving dream-images or hallucinations. This is a point that is likely to be forgotten and that, being crucial, will be repeated more than once.

Given the fact that we have mutually discrepant sense-perceptions, there are, as a matter of logic, only two paths we can take.

Path #1: We can say that *some* of those perceptions are veridical while others aren't. (It must be kept firmly in mind that, in this context, we're

setting aside patently deviant perceptions, e.g., my goblin-perception.)

Path #2: We can say that *all* our perceptions are correct. Let's discuss these two paths.

Path #1 evaluated: Supposing that we decide that some of our perceptions are wrong, it must be asked: on what basis are we making this determination?

If we do it on the basis of some principle known to us in some *non*-sensory manner, we're not empiricists any more. So, given that we're empiricists, it cannot be on the basis of such a principle.

So it must be on the basis of some principle known to us through observation—supposing that it's made on the basis of any principle at all. But all of the perceptions we're talking about are equally credible. Remember that we're setting aside dream-images, hallucinations, etc. We're focusing *only* on garden-variety perceptions—those that, unless our senses are constantly deceiving us, must be correct.

This entails, tautologically, that there is no *observational*-basis for deciding that some of the perceptions in question are to be thrown out while others are to be kept on. In other words, there is no observation-based principle on the basis of which we can decide which perceptions to keep on. And, since we're empiricists, this means that there is no principle *at all* on the basis of which we decide which perceptions to keep on.

So, given that we're following Path #1, we must *arbitrarily* choose to throw out certain perceptions while keeping others. This means that, even if we made the right choices, we wouldn't know it, since those choices weren't justified. Which in turn means that nothing we believed on the basis of sense-perception would be justified. Which, in *its* turn, means that we wouldn't know *anything* on the basis of observation.

(And, since, by assumption, we're skeptics, we can't say this.)[243]

Plus, that procedure, if adopted, would have the consequence that our senses tell us little, if anything, about the external world.[244] Since any given perception is discrepant with many others, many of our perceptions, perhaps all of them, would get the axe.

Thus, Path #1 isn't viable. So we must take Path #2. In other words, we must say that (setting aside patently deviate (mis)perceptions), our perceptions are correct.

Path 2 Evaluated: Assuming Path #2 the right one, we are stuck with a serious problem. When I look at it from one angle, the table appears square-shaped. (Let P_1 be this perception.) When I look at it from another angle, the table appears diamond-shaped. (Let P_2 be this perception.) P_1 and P_2 cannot both be accurate *if* they're both taken to concern some trans-perceptual object. A given object cannot simultaneously be, and fail to be, equiangular. (Given any property F, nothing x can, at a given time, both have and not have x.) But even though, when taken to concern some transperceptual object, P_1 and P_2 are inconsistent with each other, P_1 *per se* is perfectly capable of co-existing with P_2 *per se.* They *do* co-exist. So they can co-exist. (Two conflicting statements can co-exist.) So instead of saying that P_1 and P_2 describe some trans-perceptual object, let's say that they *are* the object that, so we wrongly thought, they describe. (Let's say, in other words, that they're constituents of that object.) Let's say, in general, that our perceptions *are* their objects (i.e., that any given perception is a constituent of its object and that any given perception is wholly constituted by some multiplicity of perceptions).

By taking this position, we accommodate the fact that we acquire knowledge through sense-perception. We know our perceptions. (Even Cartesian skeptics concede this.) If we identify our perceptions with their objects, then we know those objects—and we know them by virtue of having sense-perceptions. So this solution accommodates the presumption that we learn about objects through sense-perception. And it also makes it unnecessary to posit ***surds,*** such as equiangular objects that aren't equiangular.

6.1.1 A systematic presentation of *Berkeley's argument* [245]

Step 1: If our senses are to be trusted, properties are observer dependent. (If

you're an elephant, people appear small. If you're an ant, they appear big. If you just came out of a sauna, room temp seems cold. If you just got out of an ice bath, it seems hot.)

Step 2: The external world is totally unknowable *unless* we suppose that, typically, things are as they appear.

Commentary on Step 2: This is a corollary of empiricism. Empiricism says: what we know, we know strictly on the basis of our senses. So if our senses tell us that x is big, then we're entitled to believe x is big—period. One cannot, in general, legitimately second-guess what one's senses are telling one (at least not if one is an empiricist). One *can* throw out patently deviant perceptions (e.g., perceptions of suddenly appearing pink elephants). But this covers only an infinitesimal minority of our sensory experience.

Step 3: Things *are* knowable. It would be absurd to say that we don't know anything about rocks, trees, etc.

Step 4: Things can't have incompatible properties. A number can't be both even *and* odd. A thing can't be both big *and* small. A body of water can't be both hot and cold.

Step 5: Suppose, if only for argument's sake, that things exist independently of our perceptions of them. Suppose that, whether or not somebody perceives it, the vase exists (the same being true of the rock, the tree, etc.). In that case, things will have *incompatible properties.*

Explanation of Step 5: Remember Step 3—we *do* know about external objects. They are *not* unknown. So they *are* as they appear. Well, they *appear* to have incompatible properties. One and the same thing *appears* both big and small, hot and cold, fast and slow, etc.

Step 6 (this step is crucial to understanding Step 7, so bear with me): There's nothing incoherent in the idea that one and the same thing should *appear* both big and small, hot and cold.

Clarification: Suppose I say "Bill is in Virginia right now" and I then say "Bill is not in Virginia right now." Obviously there cannot be some one thing

of which both statements are correct representations. But there's nothing incoherent in the idea that both statements should be *made.* While it is incoherent to hold there is some one thing which they both correctly *describe,* it's obviously perfectly ok to hold that both statements (however badly they may fit together) were affirmed. Similarly, there is nothing incoherent in the idea that, at the same time, there should occur a *perception* to the effect that x has one temperature and also a perception to the effect that x has some other temperature. What is incoherent is the idea that there should be some thing of which both perceptions are correct *representations.*

Step 7: If we say that our perceptions are *representations* of objects that lie on the other side of them, then (assuming, as we are, that our perceptions are accurate) things have incompatible properties. But if we say that our perceptions are *not* representations of such objects—if, instead, we say that perceptions *are* those objects—then things *don't* have incompatible properties.

Clarification: Remember Step 6. It isn't possible for there to be two correct but incompatible *representations* of some one thing, even though it's obviously possible for two such representations to *exist.* Our perceptions give us incompatible messages (they say that one and the same thing is both big and small, is moving both quickly and slowly, etc.). So it isn't possible for our perceptions to be correct *representations* of things. But it *is* possible for our perceptions to exist. It is irrelevant that they are incoherent if taken to be representations of trans-perceptual entities.

Step 8: It follows from Step 7 that our perceptions can't be *representations* of anything.

Step 9: But remember Step 2—we *do* know about things. We *do* know their sizes, temperatures, etc. The way to make this all work is to say: things *are* our perceptions of them. We know our perceptions obviously. (So if we say that rocks, trees, etc., *are* our perceptions, then we can explain how we can know them.) And we *also* know that perceptions can't be representations of things. So we say that perceptions don't *represent* reality: they *are* reality.

Conclusion: For something to exist *s i* for it to be perceived. The perception doesn't represent the thing. The perception *constitutes* the thing.

6.2 The problem with Berkeley's argument: its failure to take into account the relational nature of perceptual information

Berkeley's argument fails. But the reason why it does must be understood in terms of a rather subtle point about sense-perception—a point that will initially be hard to understand but that, after being stated abstractly, will be made clear with some examples.

Our senses give us *relational* information. They tell us how things are in relation to one another and in relation to ourselves. They apprise us, not of absolute, but of relative heights, distances, velocities, temperatures, and masses; our eyes tell us, not that the car is going 35 mph, but that it is going faster than that other car; our muscles, when we exert ourselves, tell us not that the barbell weighs 215 pounds, but that it weighs more than the barbells we tried to move a moment ago. Our senses tell us, not that Smith is 7-feet tall, but that he is much taller than Jones. And our senses tell us where things are, not in relation to some absolute coordinate system, but in relation to us. So the information given to us by our senses is relational in two senses: objects are related to one another; and they are also related to an egocentric, as opposed to universal, coordinate system.

Once this is taken into account, it is clear why Berkeley's argument crumbles. Consider a situation where X is walking at a rate of 1 mph, Y is sprinting at a rate of 30 mph, and Z is driving at a rate of 90 mph. (X, Y, and Z can see each other; they have good vision and normal cognitive capacities. So they know, at least approximately, what their relative velocities are.) According to Berkeley, X's perception of Y is simply this: Y IS MOVING SLOWLY. And, according to Berkeley, Z's perception of Y is simply: Y IS MOVING SLOWLY. And, on this basis, Berkeley concludes that Y's state of motion (his velocity and, by similar arguments, his trajectory) exists only in the mind (or, more accurately, in the minds of X and Z and any other observers of Y). But Berkeley has misreported what the relevant individuals are being told by their own eyes. X's eyes don't tell him: Y IS MOVING

FAST. They tell him: Y IS MOVING FASTER THAN I AM. And Z's eyes don't tell him: Y IS MOVING SLOWLY. They tell him: Y IS MOVING MORE SLOWLY THAN I AM.

So when we consider what Y's eyes are telling him, and put it together with what Z's are telling *him* (Z), what we get is this:

(i) Z is moving fast then Y, and Y is moving faster than X.

But there is nothing incoherent about (i). It does indeed attribute two properties to Y (namely, the property of moving slower than Z and the property of moving faster than X). But there is no reason why a thing can't simultaneously have both properties; for those properties are not incompatible—they aren't like the properties "square" and "circle." (i) describes a state of affairs that can, without any incoherence or self-contradiction, be attributed to the external world.

This reasoning is easily extended to undermine Berkeley's argument in its entirety, as another story will help make clear. X is a tiny person—he's 1-inch tall (but otherwise like us: good vision, good reasoning skills, etc.) Y is 1-foot tall (same qualification). And Z is 6-feet tall (same qualification). They can all see one another. According to Berkeley, here's what's going on: Z's eyes tell him: Y IS TINY. X's eyes tell *him* (X): Y IS HUGE. Berkeley concludes Y is both tiny and huge and that, since nothing can be both tiny and huge, size must be in the mind of the observer and not in the thing observed. But Berkeley is wrong about what X and Z are told by their senses. X's eyes tell him: Y IS MUCH LARGER THAN I AM. Z's senses tell *him* (Z): Y is much smaller than I am. So when we put together what X and Z are told by their senses, what we get is:

(ii) X is smaller than Y, and Y is smaller than Z.

And there is nothing incoherent about (ii). It doesn't ascribe incompatible properties to Y, and it doesn't therefore require that we abandon the view that size is in things themselves, as opposed to the minds of their observers.

Some additional illustrations of these principles may be in order. X is super-strong and moves barbell B easily. Y is not so strong and can barely get B to budge. According to Berkeley, X's kinesthetic sensations (i.e., the feelings of resistance that he has when moving B) tell him: B IS LIGHT.

And, according to Berkeley, Y's kinesthetic sensations tell him: B IS HEAVY. But, contrary to what Berkeley holds, X's sensations don't give him a message of the form: B's MASS IS SUCH AND SUCH. They give him *relational* message. They tell him what the *relation* is between (a) X's making a certain degree of effort and (b) B's moving by a certain amount. So X's sensations tell him how B's resistance to motion *compares* with his own ability to overcome such resistance. By similar reasoning, Y's sensations tell him (Y) how B's resistance to motion compares with *his* (B's) ability to overcome such resistance. So when we put what X's sensations are telling him together with what Y's sensations are telling *him,* what we get is:

(iii) The ratio of X's ability to move things to B's resistance to being moved is higher than the ratio of Y's ability to move things to B's resistance to being moved.

(iii) is a perfectly consistent message. In fact, it's consistent with, and confirms, what we already knew. We knew that X was strong, and that Y was not so strong; and, given that information, it was to be expected that X would have an easier time moving B than Y and, also, that this would be borne out by the kinesthetic sensations experienced by these individuals. To sum up, once it is made clear that our sensations and perceptions give us *relational* information, it turns out that, although our senses are obviously wrong some of the time, they *don't,* contrary to Berkeley argued, *systematically* give us contradictory reports.

Here is one example to drive these points home. X is a hand that has just been immersed in ice-cold water. Y is a hand that is been soaking in water that is very hot (but not so hot that it damages Y). B is a bucket of room-temperature water. Here is what Berkeley says (this is a paraphrase, not a quotation):

"When X is immersed in B, his hand says: THIS WATER IS HOT. And when Y is immersed in B, his hand says: THIS WATER IS COLD. So X and Y say opposed things about the water in B. Therefore, B's temperature isn't in B; it's in the observer's mind."

But Berkeley is wrong about what X and Y say. Remember that perceptual

information is relational. When immersed in B, X doesn't say: THIS WATER IS HOT. It says: THIS WATER IS MUCH *HOTTER* THAN THE WATER I WAS JUST IN. And, when immersed in B, Y doesn't say: THIS WATER IS COLD. It says: THIS WATER IS MUCH COLDER THAN THE WATER I WAS JUST IN.

So when we put together the messages that X and Y present to their own, we get:

(iv) The water in the bucket is warmer than the water that my left hand was in prior to being in the bucket, and the water in the bucket is colder than the water that my right hand was in prior to being in the bucket.

There is nothing incoherent about (iv); it doesn't attribute incompatible properties to anything; nor, therefore, does it require us to deny the mind-independence of any aspect of reality.

6.3 How we actually learn what properties things have

As we've noted, our senses give us only relational information; they tell us, not that the object x weighs 80 pounds, but that it weighs *more* than object y; not that object x is moving at a rate of 50 mph, but that it is moving fast*er* than y; not that x has a temperature of 78°, but that x is warm*er* than y; and so on. At the same time, we're obviously capable of learning objects' actual masses, temperatures, and such. But how is this possible if our senses give us only relational information?

To answer this question, we must take a brief detour through the philosophy of science. Suppose you decide to measure time in terms of your own heart rate. You'll find that, when you do vigorous exercise, everything in the universe (other than your heart rate itself) slows down. (Your heart rate does not itself slow down; for, being the standard of comparison, its rate is by definition constant.) And you'll find that, when you are calm and relaxed, everything in the universe speeds up.

Here you can say one of two things. (i) Because of some strange law, your exercising causes the entire universe (except for your heart itself) to

slow down (and your ceasing to exercise causes it to speed up). (ii) You shouldn't measure rates of change with respect to your heart rate, since doing so creates otherwise absent causal anomalies and demands that relatively simple physical laws be replaced with exceedingly complicated ones.

Obviously (ii) is the right choice. You shouldn't regard your heart rate as a periodic process. You should say that it's your heart rate that's changing, not the entire universe.[246]

Comparable points hold with respect to all degree properties. (By a "degree property," I mean one that can be had to varying degrees. So coldness is a degree property, since one thing can be cold*er* than another. But the property of being divisible by two is not, since a number either has that property or it doesn't.) Suppose a consequence of taking object L to have a constant length is that, when L is put in a freezer, everything in the universe (except for L itself) expands and that, when x is put in boiling water, everything in the universe shrinks. In that case, obviously, you shouldn't take L to have an unchanging length; you shouldn't take it as your standard of comparison. The principle implicit in all this is that, if one learns that there has been a change in the difference between x and y in respect of how much they have degree property phi, one should regard the extent to which x has phi as having changed *less* than the degree to which y has phi *if and only if* doing so is less costly, from an explanatory viewpoint, than it would be to regard y as having undergone less change (in that respect) than y or to regard x and y as having undergone equal amounts of change (same qualification). So, if fewer causal anomalies are created by regarding y as having undergone less phi-change than x than are created by regarding x as having undergone more phi-change than y, then one should *not* regard x as approximating, to a higher degree than y, to a condition of phi-uniformity. And if fewer causal anomalies are created by regarding x and y as having undergone equal amounts of phi-change than are created by supposing one of them to have undergone more phi-change than the other, then one should regard *neither* x *nor* y as approximating a condition of phi-uniformity to a higher degree than the other. *Otherwise* one should regard x as approximating to a condition of phi-uniformity to a higher degree than x.

6.3.1 How this last point relates to sense-perception

Our senses constantly give us information about comparative rates of change. One's eyes might tell one, for example, that the apparent size of the man is changing faster than the apparent size of the tree. The man's apparent size, so one's eyes tell one, is shrinking faster than the tree's apparent size, even though the apparent size of each one is shrinking. Given this raw, pre-inferential, visual information, one holds any one of the following:

(i) One's distance from the man and the tree remains constant; but the man and the tree are both shrinking, and the man is shrinking faster than the tree.

(ii) One is moving *away* from both the man and the tree (and, therefore, one's distance from the man and the tree is not constant), but the distance between the man and the tree remains constant; and, whereas the man is shrinking, the tree's size remains constant.

(iii) One is moving away from both the man and the tree, but the distance between the man and the tree remains constant; and, even though *both* the man and the tree are shrinking, the man is doing so at a faster rate than the tree.

(iv) The tree's size remains constant, and so does the man's; the man is moving away from the tree in one direction, and you are moving away from the tree in the opposite direction.

Each of (i)–(iv) is consistent with the raw information contained in the previously described perception. But a world where any one of (i)–(iii) was correct would *ceteris paribus* be a world that was replete with causal anomalies absent from a world where (iv) was correct. This means that, if one were to countenance any one of (i)–(iii), one's understanding of the world and the forces that govern it would undergo a weakening that it would not undergo were one instead to countenance (iv). It is for this reason, then, that *ceteris paribus* one countenances (iv), and not any one of (i)–(iii). It is for this reason, that, after exercising vigorously, one believes that one's heart rate sped up and not that the rest of the world slowed down; that, after lifting weights for a year, one became stronger and not that everything became lighter; that, on leaving a steam room, one's body that has become warmer

and not that the world that has become colder; and so on.

So although it is obviously *through* sense-perception that we learn objects' masses, speeds, temperatures, and so on, we don't *sense-perceive* them, at least not directly. Although it is through sense-perception that one learns that the water is 78°, one doesn't *sense-perceive* that the water has that temperature. What one sense-perceives is that the water is warm*er* than x and cold*er* than y. Given this information, one is justified in believing that the water's temperature is 78° just in case, in doing so, one is undermining the integrity of one's understanding of physical reality less than one would be in not doing so. Basically, objects' actual properties are inferred from sense-perceptions and not directly sense-perceived; and the relevant rule of inference is: "of the various ways of accommodating the raw, sensory data that you've been given (and there will always be several such ways), the right one is the one that is the least *ad hoc* (that posits the fewest causal anomalies, that has the most explanatory fertility, etc.)."

6.4 Primary vs. Secondary properties

Like John Locke, Berkeley distinguishes between ***primary properties*** and ***secondary properties.***[247] Examples of primary properties are shape, size, density, liquidity, solidity, gaseousness, and state of motion. Examples of secondary properties are sweetness, smell, color, and timbre.[248]

For a thing to have a given secondary property—e.g., for it to have the taste of Snicker's bar—is for that thing to be prone to induce sensations of a certain kind in creatures of a given kind.

For a thing to have a given *primary* property is for it to have an effect of a given kind on the conditions that objects must meet to occupy a certain region of space-time. By virtue of being round (or gaseous), a thing has certain effects on would-be occupants of its current location that, other things being equal, it wouldn't have were it square-shaped (or solid) instead.

According to some, the essence of Berkeley's argument for idealism is that primary properties are really secondary properties.[249] Thus interpreted, Berkeley's argument is as follows:

> Berkeley's argument (BA): For a thing to have a given secondary property is for it to taste, smell, feel, etc., a certain way. Some philosophers hold that, in this respect, shape, size, and other so-called primary properties are fundamentally different from secondary properties. But they're wrong. Shape, size, etc., are quite as observer-dependent as taste, color, etc. When you're near the house, it looks big (we're assuming that your perception is accurate); and when you're far from the house, it looks small (same qualification). Of course, we *could* say that the size an object *seems* to have is different from the size it *actually* has. But in that case, we'd be stuck saying that our senses tell us nothing, or next to nothing, about the external world, which is absurd. It follows (for the reasons given in 6.1) that the objects of sense-perception are identical with our perceptions of them.

The problem with BA is that, contrary to what it says, your first house-perception *is* consistent with your second house-perception. The first perception doesn't say that the house is *big.* It tells you how its size compares with those of various objects. It tells you that it's shorter than the telephone pole next to it, but taller than the cactus in front of it, etc. The comparative information given to you by your second perception is consistent with it. The relative apparent sizes of the cactus, house, and telephone pole are the same.

In the second perception, the house appears the way that small things appear *when looked at from up close.* And *that* is the one and only sense in which the house "looks small." But you're not looking at the house from up close. You're looking at it from far away. And your perception isn't telling you that the house is close by and looks that way. It's telling you that you're far away from the house and it looks that way. So it isn't telling you that the house is small. It's actually telling you that it's quite big.

When you're near the house, your house-image hogs up your entire field of vision. It occupies more space on your visual map. When you're far from the house, your field of vision has to include a great many objects *other* than the house. So it doesn't eat up very much map-space. Thus, measured in terms of how much map-space they eat up, the first house-*image* is larger than the second house- *image.*

But that doesn't mean that those perceptions make different statements concerning the size of the house. On a map the size of a postage stamp, the discoloration that represents Denmark will be miniscule. On a map the size of the side of a big house, the corresponding discoloration will be much bigger. But that doesn't mean that Denmark is small according to the small map and big according to the big one. Each of those maps tells you that Denmark's actual size corresponds to the *relative,* not the *absolute,* size of the picture of Denmark on that map. The bigger that discoloration is *in relation* to the other nation-representations on that map, the bigger the size thereby attributed to Denmark. Thus, the size attributed to Denmark by the small map might be twice that attributed to it by the big map.

The same thing *mutatis mutandis* is true of visual images. A big house no more "looks small" when looked at from far away is identical than Denmark "looks small" on the tiny map. When the house is looked at from far away, the corresponding house-*image* is small (i.e., it eats up a comparatively small amount of visual space). But that's consistent with that very perception's *representing* that very house as being huge.

6.4.1 BA embodies a spurious understanding of what secondary properties are

The health effects of eating a foul smelling cut of meat are very different from the health-effects of eating a wholesome smelling cut of meat that is otherwise observationally just like the first. The way a piece of meat smells is an indication of its ability to nourish. For this reason, otherwise hard to obtain causal knowledge is easily obtained and otherwise hard to make predictions are easily made.

Differences secondary properties *always* correlate with causal differences. In many cases, day-to-day experience makes it clear what causal

properties are associated with possession of a given secondary property. (Day-to-day experience makes it clear that things with certain tastes and smells are not fit to drink.) In other case, day-to-day experience doesn't do this. (Day-to-day experience doesn't make it clear how an object that is red differs in respect of its causal properties from one that is green but is otherwise identical.) But if a given object has a given secondary property, it *ipso facto* has some causal property that it wouldn't otherwise have.

For this reason, secondary properties are not in any relevant sense "subjective." The term "subjective," as John Searle (1992) points out, has two very different meanings. To say that such and such is "subjective" can mean that

(i) its existence presupposes that of some subject,

or it can mean that

(ii) it embodies prejudices or other subject-specific idiosycracies that tend to inhibit the acquisition of knowledge.

My belief that 2 + 2 = 4 is subjective in sense (i). That belief presupposes the existence of a subject. There can't be disembodied beliefs floating about. [250] But that belief isn't subjective in sense (ii). It is (so we will assume) based on rational consideration of the relevant information. By contrast, my intense fear of people who wear plaid shirts (which is based, not on rational consideration of the relevant data, but on childhood traumas that we needn't discuss) is subjective in sense (ii). In other words, it is rooted in peculiarities of my own person that make it harder for me than it would otherwise be to see things as they are. My olfactory perception that the maggot-infested meat in front of me smells awful is subjective in sense (i) but not in sense (ii). It is subjective only in the trivial and irrelevant sense in which *any* mental entity, even a coolly made judgment, is subjective: it is subjective in the same way as my rational and informed belief that 2+2=4 is subjective. But it isn't subjective in the same way as my prejudice-driven fear of people who wear plaid shirts.

Bearing this in mind, suppose that, all of a sudden, the way things smelled was no indication as to whether they were rotten or not. So, for example, a putrid-smelling cut of meat might or might not be good to eat—its

smell would give you no information as to its fitness as a source of nourishment—the same being true of a wholesome-smelling cut of meat. In general, we wouldn't know anything about an object's causal properties from the way it smelled. But we often know a lot about a thing's causal properties from the way it smells. Smells correlate *extremely* reliably with certain causal features of objects—sometimes much more reliably than other properties. Sometimes the way thing smells is the only indication (short of eating it or sending it to a toxicology lab) that it's laced with poison or that it's completely rotten.

Smells track primary properties; they track facts about microstructure that would otherwise be hard to detect. If a police-detective suspects that X is pure cocaine, he can verify his hypothesis by (i) snorting X, which would have various adverse consequences, (ii) sending X to a forensics laboratory, which would be extremely slow, or (iii) tasting it, which would be quick and easy.

A relevant fact about method (ii) is that many scientific techniques would have been impossible to develop had there not been reliable correlations between microstructure and odor. This is true of secondary properties generally.

There are some *apparent* counterexamples to our contention that secondary properties track microstructure. For example, a thing's color doesn't correlate reliably with its microstructure. From far away, the ocean water looks black; from up close it's transparent. Locke concluded from this that secondary properties are subjective.

But this is not the only conclusion to draw. An object's apparent color is a function *both* of that object's microstructure *and* of the relevant observer's physical relation to that object. But so does an object's apparent shape, apparent size, smell, etc. So the variability of apparent color no more entails the subjectivity of color than does the variability of apparent shape. And so long as the conditions of observation are held constant, an object's apparent color does reliably correlate with its microstructure, just like its apparent shape (though not, it must be admitted, to as high a degree).

Our *perceptions* of primary properties are very different from our perceptions of secondary properties. When one sees a porcupine, one's visual image has a structural similarity to that of the porcupine; given any two

features of the porcupine that are distinguished *by* image, there are corresponding differentiations *in* that image.[251] Perceptions of secondary properties are very different; they *aren't* internally differentiated, at least not in a way that at all corresponds to the objective facts to which they are hewed.

This, presumably, is at least part of the reason why it's widely held that an object's secondary properties are really properties of the subject's experiences of that object, and not of that object itself. (It's also why it seems more natural to speak of "sensations" of taste, smell, etc. than of "perceptions" of them.) But, I will now argue, an object's secondary properties belong to it no less than its primary properties. I must warn that the forthcoming argument is quite intricate and that, for the first few paragraphs, it will seem as though I'm arguing for the contention that an object's secondary properties *don't* really belong to it. I must ask for a certain amount of patience on the reader's part.

Let S be the smell that, in actuality, is had by extremely putrid meat; and let S* be the extremely pleasant fragrance that, in actuality, is given off by eucalyptus trees. There is no *inherent* reason why eucalyptus trees couldn't have S and there is no inherent reason why rotten meat couldn't have S*. In other words, if there occurred a smell-exchange between eucalyptus trees and rotten meat, we would not, at least not for that reason alone, forfeit the causal knowledge that we now have by virtue of knowing that a cut of meat has S or that a eucalyptus tree has S*.

Of course, there would *initially* be a loss of knowledge. One would mistakenly think that the pleasant-smelling cut of meat one was about to eat was not rotten. But given some time to adjust to the new smell-situation, one would cease to make such mistakes.[252]

There is another reason why, even though it wouldn't have any lasting affect in our ability to acquire knowledge through olfaction, that smell-exchange would initially compromise our ability to do so. This fact *initially* suggests that olfaction doesn't give us knowledge of the causal structure of the world. But, duly scrutinized, it turns out to entail that olfaction is, like vision and audition, *does* give us such knowledge.

Much of the knowledge that we acquire through the sense of smell is based on the comparative properties of our olfactory sensations. Other things being equal, the more rotten a cut of meat is, the worse it smells. Other things being equal, the more pungent the fragrance of a eucalyptus tree is, the richer that tree is in the chemical responsible for that smell.

Bearing this in mind, suppose that S is the exact smell had (under a given set of circumstances) by a cut of meet that was rotten to degree n, and that S* is the fragrance (same qualification) of a eucalyptus having amount m of the relevant chemical. Because the previously described exchange hasn't occurred, given a cut of meat having S and some other, otherwise observationally identical cut of meat, with a much stronger (weaker) version of S, I know on that basis that *ceteris paribus* the latter cut of meat is more (less) rotten than the former. For this reason, if the exchanged in question occurred *but that other, compensatory changes didn't occur alongside t i ,* our ability to acquire comparative knowledge of this kind would be diminished.

But there's no reason why any given set of causal properties should be associated with any given type of olfactory sensation. What *is* necessary is that, once the assignments are made, their mutual relations parallel the mutual relations of their objects.

> (SA[253]) So you admit that its having this as opposed to that odor does not, in and of itself, diminish our ability to learn about the world through the sense of smell. You admit, in other words, that rotten meat doesn't *have* to have the smell it does. But in admitting this, you're admitting that it doesn't *really* have that smell.

There are two problems with SA.

Problem #1: When we say that a given cut of meat has a bad smell, *we're talking about the cut of meat.* If the word "smell" referred to *sensations,* then sentences like "that meat has an awful smell" wouldn't be about the meat. By the same token, so far as that sentence *is* about the meat, it isn't about sensations. Thus, to the extent that it really is about the meat, whether or not that sentence makes a true statement has nothing to do with anyone's sensations. And to the extent that it isn't about our sensations, "that meat has an awful smell" is about the meat's causal properties or, in any case, is about

some other (probably related) mind-independent fact.

Problem #2: SA establishes that smell is mind-dependent *only* to the extent that it establishes the obviously, or at least presumably, false conclusion that size and shape are mind-dependent.

Look at the book in front of you. Let V be the visual sensation you are having. Under otherwise unchanged circumstances, you could have a visual sensation that, although very different from V, gave you either the very same information or that gave you much more information. In human beings, the set of sight isn't nearly as acute as it is where certain other species are concerned. Were they to look at the book in front of you, under conditions just like the ones that actually obtain, they'd have sensations very different from V that contained all the information contained in V, and then some. [254] You could have visual sensations very different from the ones you actually have without, for that reason, failing to know anything that you do know.

But this doesn't entail that the facts of which vision apprises us are mind-dependent. V gives you a lot of information about the book's location, its shape, and its size. And V's shape, size, etc., are quite definitely mind-*in*dependent.

"But aren't you begging the question against Berkeley in saying this?," it will be asked. "Wasn't Berkeley arguing that, since (as you just pointed out) secondary properties are relevantly similar primary properties, to be is to be perceived?"

Berkeley does indeed argue that. But he also takes it for granted that secondary properties are mind-dependent. In other words, his argument has the form: (i) secondary properties are mind-dependent; (ii) secondary properties are relevantly similar to primary properties; (iii) therefore, primary properties are mind-dependent; (iv) it follows from (iii) that objects are mind-dependent—are creatures of the mind.

But I argued that secondary properties *aren't* mind-dependent. I reject step (i), and provided an independently plausible reason for doing so. Therefore, what I've said doesn't beg any questions against Berkeley.

6.4.2 Is temperature a primary or a secondary

property?

Some properties that seem to be secondary turn out to be primary. A brief examination of one such property will clarify some of the points just made.

For an object to have a certain *temperature* is for the particles composing it to have a certain mean kinetic energy. (Basically, it's for those particles to move around in a certain way. The faster the movements, the greater the temperature.[255]) But it's initially tempting to categorize temperature with sweetness and, thus, to see it as secondary property.

There are a couple of reasons for this. First, secondary properties are modality-specific. Sweetness can be tasted, but not heard, seen, touched.

By contrast, primary properties are transmodal. Squareness can be seen or touched. In fact, creatures with very good hearing (e.g. bats) can *hear* shapes, just as we can see them; bats can detect an object's shape through sonar. And creatures with a very olfactory sense (e.g., moles) can *smell* shape.

According to some, including Locke and Berkeley himself, temperature can be known only through touch. But this isn't true, of course. Given what temperature is (namely, molecular motion), it is just as capable of being known through sight or audition as it is of being known through touch. But it's obvious why someone would think otherwise.

There's another reason why temperature is often falsely regarded as a secondary property. There isn't anything inherently pleasant or unpleasant about seeing a square-shaped object or a fast-moving object. Under certain circumstances, it may be very unpleasant to see something move with a certain speed. But the connection is circumstance-specific. There is, however, something inherently unpleasant about certain odors, tastes, and timbres. In general, primary properties are not inherently associated with aesthetic properties—any such association is circumstantial—whereas secondary properties *are* non-circumstantially associated with such properties. Very high and very low temperatures are non-circumstantially associated with unpleasant sensations, and moderate temperatures are non-circumstantially associated with pleasant ones.

Nonetheless, temperature is a primary property, since an instance of

temperature is identical with various instances of motion.

6.5 A second Berkeleyan argument—"to be is to be conceived"

At one point, Berkeley argues that to be is to be conceived ("esse concipi ist"). In other words, things are the awarenesses that we, or other thinking beings, have of them.

Here's his argument. Try to conceive of something that exists but isn't conceived. You can't do it. Why not? Because you just conceived of it. If you try to conceive of some chicken that isn't conceived of, then you just conceived of it, and it isn't unconceived. Thus, nothing unconceived can be conceived to exist. Therefore, to exist is to be conceived.

The problem with this argument is that, contrary to what Berkeley assumes, *propositions,* not *objects,* are the objects of conception. When one "conceives of a 50-pound chicken," what one is doing is considering a proposition, viz: *there exists a chicken that weighs 50 pounds,* which (see Chapter 1, Sections 2.0–2.4) is equivalent with *the property of being a 50-pound chicken* is instantiated. There isn't some specific chicken x such that x weighs 50 pounds and one is thinking about x. When one conceives of an unconceived chicken, there isn't some specific chicken x that one is thinking about but that, at the same time, no one is thinking about. What's going is that one is considering the proposition: *there exists some chicken or other that nobody is thinking of,* which is equivalent with the proposition: *the property of being a 50-pound chicken that nobody is thinking of is instantiated.* And that proposition is clearly not an absurd one. So Berkeley's argument fails.

7.0 The phenomenal reduction—some preliminaries

First some terminological points. An ***object-statement*** is one that concerns mind-independent entities, e.g., "that statue is over ten feet tall" is an object statement. A ***perception-statement*** is one that concerns one's own perceptions (e.g "I am having an experience that, if it were veridical, would be of a statue that is over ten feet tall"). Two statements are ***observationally equivalent*** if there is no observation that (dis)confirms the one either more or

less than it (dis)confirms the other. (All analytically equivalent sentences are observationally equivalent—but not vice versa, as we're about to see.) Finally, a property or relation is ***observationally inert*** if no instance of it could possibly affect what anyone observes.

7.1 The phenomenal reduction

For the reasons given earlier, strict empiricism leads to either skepticism or to idealism. Skepticism and idealism are both deeply at odds with commonsense. Berkeley agreed that skepticism was at odds with commonsense. But he denied that idealism had this defect. He held, in fact, that idealism *is* the commonsense view.

In an effort to vindicate this obviously false claim, Berkeley argued that any object-statement that we could have any good reason to believe correct can be reinterpreted as a perception-statement.

Here is a modernized presentation of Berkeley's reasoning:

(BR[256]) A statement about the external world is meaningful only to the extent that, were it true, we'd have observations that we wouldn't otherwise have.

This entails that such a statement is meaningful only *to the extent* that it concerns observations. When the possible observations run out, so does the meaning. Thus, "Smith has a car" is without meaning to the extent Smith's having a car is observationally inert, the same thing *mutatis mutandis* obviously being true of "Brown has a car." And

(1) "Smith's car weighs twice as much as Brown's car"

is meaningful only to the extent that, for any objects x and y, if x weighs twice as much as y, there would *ipso facto* occur observations that would not otherwise occur.

If we replace all of the constants in (1) with variables, what results is the sentence-form:

(2) A bears R to B.

A true sentence results if the variables in (2) are replaced with constants referring to classes of perceptions. Let K_1 be the class of perceptions that one has in virtue of the fact that Smith's car exists. (K_1 might include a perception of Smith pulling out of his garage.) Let K_2 be the class of perceptions that one has when, and only when, there are objects x and y such that x weighs twice as much as y. And let K_3 be the class of perceptions that one has in virtue of the fact that Brown's car exists.

If a given observation O confirms (1) to a degree, it confirms

(3) K_1 bears K_2 to K_3

to that same degree. (In this context, "confirms" is to be taken to refer to positive or negative confirmation.) If a given observation O* (dis)confirms any one of these three claims, it is *ipso facto* (not) a member of one of K_1–K_3. So there can't be an observation that (dis)confirms (1) more or less than it (dis)confirms (3).

There is therefore no observation-based reason to prefer (1) to (3). So assuming, as we are, that empiricism is correct, and that all good reasons are observation-based, it follows that there is no reason at all.

But *the very fact that there is no such reason* is itself a reason to prefer (3) to (1). (3) doesn't posit anything for which there is no observational evidence. (We'll henceforth use the word "trans-observational" to describe entities for whose existence there is no observational evidence.) But (1), at least as many philosophers interpret it, *does* posit trans-observational entities.[257] And to the extent that a sentence demands the existence of trans-observational entities, it is meaningless. Why? Because sentences are meaningful only to the extent that, if true, they make a difference; and, given the truth of empiricism, a sentence makes a difference only if it makes an *observational* difference.

7.2 The viability and significance of the phenomenal reduction

In this context, I'll use the word "perception" to denote any mental state that, *if veridical,* is an accurate sensory observation but that needn't be veridical. Thus, hallucinations are perceptions, as we'll be using that word, and so are veridical perceptions.

You have a perception of Smith driving a car. Let P be that perception. P is obviously evidence, albeit ***defeasible***[258] evidence, that Smith is driving a car. Berkeley admits this. In other words, Berkeley admits—as he surely must—that, other things being equal, your having a perception of Smith driving a car gives some weight to the contention that Smith is driving a car.

But *how* is P evidence of Smith's driving a car? It's evidence of it because it *says* that Smith is driving a car. If your mother, who you trust, calls you up and says "Smith is driving a car (which he shouldn't being doing, since he's blind drunk)," you're likely to believe that Smith is driving a car, even if you didn't previously have any reason to believe it. The reason is that you regard your mother's testimony as evidence to the effect. You also regard your visual perception of Smith driving as (extremely strong) evidence to that effect. You do this because your eyes are telling you—much as your mother might, but much more convincingly—that Smith (or, at any rate, somebody who looks a lot like him) is driving a car.

Part of what your eyes are telling you is that some spatially remote object (namely, Smith) bears a certain relation (namely, that of driving) with respect to some other spatially remote object (namely, the car Smith is driving). And in telling you that Smith and his car are spatially remote from you, your eyes are saying, or at least implying, that Smith isn't a part of you or, therefore, of your mind.

It may be that P is wrong. It may be that you're hallucinating and that, when you come down from your acid trip, it will turn out that Smith wasn't driving at the time in question. But P still *said* that he was driving, and it still *said* that there exists some spatially remote object bearing a certain relation to some other such object.

Any given perception is to the effect that objects that aren't a part of you bear certain relations to one another and/or individually have certain properties. So to the extent that our sense-perceptions are evidence of anything, they are evidence of trans-perceptual objects.[259]

And that is why statements about perceptions cannot possibly be equivalent with statements about perceptions.

8.0 Two insights of Berkeley's

Berkeley's argument for idealism does not go through, and neither does his attempt to reduce statements about objects to statements about perceptions. But, in the process of trying to establish idealism, Berkeley had two extremely profound and original insights, whose philosophical consequences are still being fathomed, to wit:

(A) many of the "statements" we make aren't really statements at all, and are instead *statement-forms,*

and

(B) a statement's grammatical form may not coincide with its logical form—that, in other words, a statement must be reparsed if its meaning is to be made clear.[260]

(A) *explained:* "The number five is odd" is a statement. It attributes a definite property to a definite object. The expression 'x is odd' isn't a statement. It *would* be a statement if the variable were replaced with a constant. A given statement form may be true 'under some interpretations of it' and false under others. This means that replacing the variables with constants may yield truths or falsehoods.

An "interpretation" of a statement-form is simply an assignment of constants to the variables. To ***model*** a statement-form is to identify an interpretation of it that yields a true statement. Here is a trivial example:

(1) A bears relationship R to B.

(2) If x bears R to y, y bears R to x.

(3) Nothing bears R to itself. In other words, ‹x bears R to x› is false, for all objects x.

Let R be the relationship that one bears to somebody else in virtue of being that person's next-door neighbor. Since (setting aside the odd person who owns adjacent houses) nobody is one's own next-door neighbor, (3)

comes out true. So does (2), since x must be y's next-door neighbor if y is x's next-door neighbor. Thus, supposing for argument's sake that Smith and Jones are next-door neighbors, (1) comes out true if "A" refers to Smith and "B" refers to Jones. We have thus modeled (1)–(3).

In conjecturing that (so-called) statements such as 'that rock weighs five lbs' can be reinterpreted as statements about perceptions, Berkeley was in effect saying that such statements aren't statements at all—they're statement forms. Though he didn't himself put it this way, he was saying (i) that the terms in them (e.g., 'rock,' 'weighs five lbs') are really variables, and (ii) that true statements would result if those variables were replaced with constants referring to perceptions.

He was wrong about (ii), as we saw a few pages ago. But he *was* right about (i). There is no one microstructure that Rover must have if my current perception of him is to be accurate. There is, as we will now see, no one way anything must be if *any* given perception is to be correct.

We've seen that our knowledge of the world is strictly comparative.[261] We know how fast a given thing is traveling *only* to the extent that we know how much fast*er* it is traveling than some other thing. We know how heavy/hot/big/etc., a thing is *only* to the extent that we know how much heavier/hotter/bigger/etc., that thing is than some other thing. Spatiotemporal knowledge is comparative knowledge.

That means that *any*thing—no matter what it's parts are 'made of'—whose parts are appropriately interrelated is no more, and no less, consistent with what our sense-perceptions tell us than is any other thing whose parts are thus interrelated. This means that perception apprises us only of *structure.*

Our perceptions (if correct) tell us the structure of the world, and our physical theories (if correct) apprise us of micro*structure.* Given *any* two things having the requisite structure, there cannot possibly be any observational basis for holding that it, and not the other thing, is what is really out there.

This means that, if x and y both have the requisite structure but are qualitatively different in some non-structural respect, the supposition that x is what is really out there models the data exactly as well—no better, no worse—than the supposition that y is what is really out there.

This doesn't mean that there is no fact as to what is out there. There obviously is. Given any set of possible objects, each having the structure of the external world but each differing in some non-structural respect from the others, at most one of those objects can be the external world. But the supposition that any given one of them is what is actual does exactly as good a job of modeling the observational data as the supposition that any other given one of them is what is actual. And one of the claims embodied in Berkeley's argument(s) for idealism is precisely that, to the extent that the probability of a hypothesis is a function of its congruence with observational data, anything that models the observational data *ipso facto* no less likely to be what's really out there than anything else that does so. Berkeley was therefore (so far as I know) the first model-theorist.

(B) explained: To my knowledge, Berkeley was the first to recognize the distinction grammatical structure (surface structure) and logical form (deep structure[262]). To avoid getting caught up in technicalities, let's also suppose Smith is the only percipient being in existence. Berkeley's idealism seems to entail that, if Smith closes his eyes (and is otherwise not in sensory contact with the world), the whole world disappears; and it also seems to entail that the whole world springs back into existence when Smith opens his eyes. This is deeply implausible.

It also seems to undermine Berkeley's claim that any object-statement that is presumed correct (e.g. 'the world doesn't' disappear when I close my eyes') is equivalent with a statement stating the observational evidence for that claim. When I close my eyes, the *observational* evidence seems to suggest that the world has vanished. And it's only, so it is thought (or so I personally think), because one knows that one's observations *aren't* the one and only relevant yardstick that one knows that the world hasn't disappeared.

Berkeley's response to this is brilliant. He says that the real meaning of

(1) "the world doesn't go out of existence when Smith (who, for argument's sake, we'll assume to be the only sentient being in existence) closes his eyes"

is ***perspicuously***[263] represented by a sentence whose surface-structure bears no resemblance to (1)'s. That other sentence is one along the lines of:

(2) “if Smith closes his eyes for x minutes and then reopens them, the things he sees are where one would *expect* them to be, *given* that they are subject to the usual forces.”

Berkeley’s own illustration of this principle is a bit more crisp than this one. The stars are moving, Berkeley points out. A more precise way of putting it, as Berkeley himself makes clear, is that we are moving relative to the stars, and the stars are therefore moving relative to us. But, Berkeley points out, we don’t *see* the stars move. The blips in the sky that we see on a given night aren’t in the same relative positions as the blips in the sky that we see on other nights.

For this reason, we do a kind of “regression analysis” on the blips that we see over the years. We try to connect the blips seen on night 1 with those seen on night 2, and we try to connect the latter with those seen on night 3, and so on. So if $N_1...N_n$ are nights in question, and n is some high number (e.g., 500), we identify curves such that, if it is assumed that those curves are the trajectories of those blips, it follows that those blips obey well-defined, relative simple laws. So the logical form of:

(A) “the stars are moving even when nobody is looking at them”

is perspicuously given by the grammatically very different sentence:

(B) “if it is assumed that, when we’re not looking at them, the stars obey certain laws and, therefore, have certain trajectories, we can make observationally accurate predictions that we couldn’t otherwise make, and arrive at observationally correct explanations that we couldn’t otherwise arrive at.

Berkeley said, as clearly as it can be said, that what sentences really mean isn’t what, given their overt forms, they seem to mean. So setting aside some petty and insubstantial objections relating to the exact meanings of “grammatical” and “logical,” Berkeley was, to my knowledge, the first person to recognize the distinction between grammatical and logical form. [264]

PART IV

Necessity, Causality, and Personal Identity

Chapter 14

Determinism, Randomness, and Unpredictability

1.0 Determinism and indeterminism

Determinism is the view that how the world is at any given point in time fixes how it will be in every respect at all later times. So given how the world is now, there is only one way it could possibly be in a thousand (or a billion) years. And given how the world was a minute ago (or a billion years ago), the world *has* to be the way it is now: things couldn't possibly be any other way.

Indeterminism is the doctrine that determinism is false. So it is the view that, for at least *one* moment in time, how the world is at that time does not completely fix how the world will be at all later times.

Indeterminism is *not* the view that the world is chaotic or lawless. Indeterminism is actually an extremely moderate doctrine. All it says is that, at *some* point in time (it could be past, present, or future), some event occurs that doesn't have to occur. So, technically speaking, for indeterminism to be true, there needs to be only one "renegade" event.

Determinism is thus much harder to establish than indeterminism. After all, the determinist says that, for *any* time t, how the world is at t determines how the world will be in *every* respect at *all* later times. Which is a very strong claim. Indeterminism, on the other hand, only says that there is *at least one* counterexample to what the determinist says. Which is not a very strong claim.

Nonetheless, even though determinism is a strong statement for which it is difficult to produce definitive grounds, we know that it is true or at least a very close approximation to the truth. It is an open question whether, at the sub-atomic level, the world is strictly deterministic. But whatever indeterminacies there are at that level seldom, if ever, percolate up to the molecular, cellular, biological, planetary, or astronomical scales. So, for most intents and purposes, the world in which *we* operate—the world of "medium sized" objects—is deterministic.

To be sure, the world often *seems* to be indeterministic. But careful

scrutiny of any such situation, with the possible exception of those at the sub-atomic level, tends to show that it is governed by deterministic mechanisms.

2.0 Determinism vs. predictability

Predictability and determinism are not the same thing. If a system is completely *in*deterministic, then it is indeed unpredictable (or only fortuitously predictable), since prediction requires regularity and there is regularity only where there are deterministic laws. But a deterministic system may be unpredictable. Even if the world is completely deterministic—even if the state of the universe at any point completely fixes how everything will be, in every respect, at all later points—there are limits to how much any creature could predict about the course of future events. It doesn't matter how intelligent that creature is or how sharp its senses are or how good its technology is. For reasons of logic, no creature is able to predict everything that happens. There are two reasons for this.

2.1 Why predictability ≠ determinism: Reason #1[265]

We learn about the external world through *sense-perception*—that is, through vision, hearing, touch, etc. Let's focus on vision for the moment. Suppose that you're seeing some rock. (Let's refer to that rock as "R.") Here's what's involved in your seeing R. Some light beam bounced off R. (We'll refer to that light beam as "L.") As a result of that collision, L's structure changed in such a way that R's structure is now coded in it's (L's) structure. For this reason, L bears information about R. L hits your eyes, disturbing them. Those disturbances reflect L's structure. Given that L reflects R's structure, it follows that these disturbances of your eyes reflect R's structure. Those ocular disturbances therefore encode information about R. The resulting neural impulses transmit this information to some part of you that decodes this information. The results of this decoding are embodied in a visual perception that succeeds in representing R.

Light is energy. Things change when bombarded by energy. They don't always change very much. But there is always *some* change. It takes time for light, or any other information-bearing process, to travel. So by the time L

reaches your eyes, some time has elapsed. During the time that has elapsed, the changes in R that were *initiated* by L led to other changes in R. Since, for the time being, L is your only source of information about R, and since L gives you dated and otherwise inaccurate information about R, the information encoded in the visual perception that L caused you to have is itself inaccurate.

Before we take the next step, I'd like to clarify this point. Let t be the time that you have the visual perception of R that L caused you to have. Obviously t will happen after the time L hits your eyes and t is later than the time that L collided with R. During the time between that collision, on the one hand, and t, on the other, R has changed. Why has it changed? Because the just-mentioned collision made it change. That collision, as we just pointed out, set off processes because of which R changed.

What you know, at t, about R is confined to what L tells you. And what R tells you is dated information—it's yesterday's news. What you know, by virtue of having the sense-perception in question, is now dead and gone. What it tells you is how R was *a little while ago;* it doesn't tell you how R is now.

Supposing that R is some rock in your backyard—and not some rock in the depths of outer space—the time lag is minimal. And, no matter where R is, L did extremely little to change it. The changes that L induced in R, by colliding with it, fall well below the limits of sensory discrimination.

But mentally replace R with something small—something so small that the photons involved in its being bombarded with light *do* appreciably change it. In *that* sort of case, the thing you see at t is *very* different from the thing that is now out in there in the world.

You might be aware of this gulf between the perception and the reality. But any attempt on your part to narrow this gulf will only make it wider. If you try to learn more about R—by, for example, shining more light on it—you'll make it *more* unlike the thing it was initially. And you'll only compound the problem if you try to undo those distortions with counter-distortions. Setting aside the problem just discussed, there's no point in trying to reverse the effects that L had on R by shining more light on R, since you don't know how R was before you bombarded it with light. Therefore, you don't know how R has changed and, consequently, you don't know what the

needed corrections are.

What is true of your seeing R is true, to some degree or other, of *anything's having any sense-perception of anything.* It doesn't matter whether the percipient beings in question are super-brilliant aliens or iguanas or God Herself. As a matter of logic, perceiving a thing, involves having an experience that is caused by that thing. And, as a matter of logic, such experience involves that thing's being disturbed by something that eventually made its way to you. Whatever changes that disturbance had on the object, you don't know about them, since they occurred *after* the relevant information-transmitting process left that thing. (Let's assume that process to be a light ray.) For the time being, that light ray is your one source of information about that thing. So whatever that light ray did to that object—whatever changes, if any, it made—you can't know what they are. If no changes occurred, you can't know it; and if changes did occur, you can't know that either, and a *fortiori* you can't know the exact nature of those changes. So what you see is different from what is out there, in the relevant place, at the time of your perception. And, as we saw a moment ago, attempts to remove this margin of error by having more perceptions will only make it bigger.

We are not talking about the limitations of human beings. We're taking about the conditions that must, as a matter of logic, be fulfilled by any sense-perception on the part of anyone or anything. If it's said that God could "just know" what is out there—that She (or He or it) didn't have to be on the receiving end of some causal process starting with external objects—I would reply by saying: "rubbish!" If there's no causal process, then, by definition, there's no information for God to go off of. And if there's no such information, then any image God comes up with is a hallucination. It might be a correct one. But, if so, it'll only be an *accidentally* correct one, and it won't be enough for knowledge of anything, and any belief She has in consequence of having that image will be nothing more than a guess. A correct guess, perhaps. Some guesses are lucky guesses. But lucky guesses aren't knowledge. (See Chapter 10 for an explanation of this.)

If it's said that the laws of logic don't apply to God, I would say: "yes they do." If, in response to that, it's said that God, being omnipotent, can't be hemmed in by anything, even the laws of logic, I would reply by saying that

this statement embodies a total failure to understand the nature of logical truth—that God's inability to create a square circle is no more an abridgment of Her powers than an athlete's "inability" to score a touchdown in a soccer match shows a lack of athletic talent. If asked why, I can only say: read Chapter 25, Section 4, which is dedicated to this problem.

In any case, the idea God could *know* of some event without having any causal connection is in the same category as the idea that a sufficiently good athlete could score a touchdown in the course of a soccer match. It's simply incoherent.

But why exactly is it incoherent? Only *justified* true beliefs are knowledge. Lucky guesses don't count. A justified belief is one for which there is evidence. One spatiotemporal thing is evidence of another if and only if the first is causally related to the second in such a way that, given the existence of the first, that of the second can be inferred. Therefore, one's own mental states are evidence of a given external thing, and therefore give one a justification for believing that thing to exist, *only* if those mental states are effects of that thing. So the idea that God could know of some spatiotemporal entity without being causally connected to that entity is no more coherent than the abject absurdity that one event could be evidence of another even though the occurrence of the first event provided no reason to suspect the occurrence of the second. Since the latter idea is obviously a non-starter, which deserves to be put in the same category as "squares don't have four sides," the same is true of the former. For similar reasons, one would simply be redefining the terms "'justification" and "knowledge" were one to say that unjustified beliefs were knowledge or that beliefs for which one had known evidence were justified. Nothing that follows from definitional truths can be false; anything that follows from a truth of logic is itself a truth of logic. So it is a truth of logic, since it follows from truths of logic, that God couldn't have knowledge of some spatiotemporal entity without being causally connected to that entity. It's also a tautology that causal connections are disruptive and that tautology entails that God's mental representations of the world must, in at least some cases, be discrepant with the corresponding realities.

Before closing the argument, let's sum up. Learning about the world involves changing it. A consequence is that the world is different from how it is represented to you as being. So even if the world is 100% deterministic,

there is necessarily a gulf between the perception and the reality.

Conclusion: Given the existence of the aforementioned gulf, any predictions one makes are necessarily slightly off. A prediction involves two things: first, knowledge of initial conditions (how things are at the moment—or, in any case, how they were just a moment ago); second, what the relevant causal laws are. Given a knowledge of those causal laws, one can predict what must occur, given those initial conditions. But since one's knowledge of initial conditions is necessarily off, though sometimes not by very much, one's predictions will be off.

2.2 Why predictability ≠ determinism: Reason #2

What we know affects how the world is. Why is this? Because what we know affects what we do, and what we do affects how the world is.

We can't know what it is that we'll know. Suppose that, right now, you know exactly what it is that you'll know in 10 years. (Let K be that body of knowledge.) In that case, what you'll know in 10 years will be K *plus* all the knowledge you pick up between now and then. And if you build *that* knowledge, viz. the knowledge you pick up along the way, into your definition of K, then you'll *already* know it and you won't pick it up along the way anymore, and the knowledge that you *do* pick up along the way will be something else. So it makes no sense to say that you could know now what it is that you'll know in 10 years.

There are trivial exceptions to this. For example, if you know that you'll be dead in 10 years, then you know that you'll know nothing. But assuming that your knowledge *grows,* you can't know how it will grow. And since how it grows affects what you'll do and therefore affects how the world will be, you can't predict how the world will be—even if, as Einstein thought, it's completely deterministic.

Thus, the very concept of an infallible self-predictor is an *incoherent* one. In other words, it makes no sense to suppose that a thing should be able to predict its own future states with complete accuracy.

2.2.1 A variation of Reason #2

There are self-fulfilling prophecies and self-defeating viewpoints. Self-

confidence breeds success. You succeed because you believe, perhaps without any good reason, that you will succeed. I've found this to be 100% true in my life. I've found that self-confidence, even (especially) in the total absence of any reason for it, always works to some degree or other. And I've found that unwarranted pessimism has some tendency to breed failure and thus to become warranted pessimism. (For some reason, unwarranted self-confidence is, in my experience, more likely to lead to success than unwarranted self-doubt is to lead to failure. But this may be an illusion; it may be that there are cases where my self-doubt was really a sham—a way of keeping self-destructive, hubris-activated forces at bay, by convincing them that I wasn't in fact guilty of excessive pride.)

The fact that self-predictions make themselves come true might initially appear more consistent with the idea that we can know the future than with the idea that we cannot. But this appearance is misleading. Self-predictions affect the future. In some cases, they do this because they are self-fulfilling. In other cases, they do it for very different reasons. In any case, given that one's predictions affect the future, so do one's predictions about the effects of one's predictions. Therefore, the self-fulfilling nature of some predictions about oneself open up a gap between how things will be and how we can know them to be.

3.0 Determinism vs. randomness

"Random" doesn't mean the same thing as "undetermined." A deterministic world isn't necessarily one in which nothing random happens; and, give or take a few niceties, an indeterministic world isn't necessarily one in which anything random does happen. A story will make it clear why these points hold.

Smith arrogantly insulted a high-ranking member of a particularly lethal criminal organization, and that person has sworn to kill him. For unrelated reasons, Smith's wife also wants him dead, and she has just hired a hit man to kill him. She has always wanted him dead; she just married him for his money. Plus, Smith has a serious heart condition, and his doctor is surprised that he didn't die years ago. You are a friend of Smith's, and you know all of this. But you don't know anything else that is relevant to an estimate of how much longer he is likely to live. Given that information, the right estimate is, of course, "not long."

But Smith beats the odds. Nobody kills him and nothing goes wrong with his heart. But one day, as he's making a sales pitch to some prospective clients, a ceiling-lamp comes loose and falls on his head, killing him.

This was a "random" occurrence. But it wasn't *uncaused.* When a ceiling lamp falls, there's a reason for it (e.g., the bolt wasn't screwed in tightly enough or the weight of the lamp was too much for the beam it was hanging from). It was "random" because, although caused, it wasn't the sort of thing that, given the information at your disposal, could reasonably be expected to do him in. Thus, it was "random" not because it was *un*caused, but because it was *independently* caused, meaning that the causal sequence that led to his death was independent of the causal sequences that were under way that *were* reasonably believed, given the information available to you, to eventuate in his demise. So, in this context, "random" doesn't mean "uncaused," it means "uncaused by anything that we had reason to believe would occur."

And this is what "random" typically means. It may be that microphysicists deal with uncaused events. But the rest of us deal exclusively with things that are quite definitely caused. (As previously stated, the indeterminacies at the micro-level don't percolate up to the macrolevel, at least not enough to engage the structures that change human lives.) So the "random" things we deal with are things that have unforeseeable causes.

A consequence is that randomness isn't a property of events; it's a relation that an event has with respect to a given body of knowledge. Relative to what you know, it is unforeseeable, and therefore random, that a ceiling lamp will kill Jones. But relative to what Howard the building inspector knows, that same event is not random. (Howard knows that the plank in question will go at any time, and he sees Smith standing underneath it.)

Thus, randomness isn't a property of events. It's a property of *statements* —or, more accurately, of relations between statements. An event is random, relative to a certain body of information, if its occurrence couldn't reasonably be predicted on the basis of that information. So randomness is really a relation that holds between the statements describing some body of data and the statement describing some event; that relation holds if the former statements don't warrant acceptance of the latter.

It's obvious that there can be random events in a deterministic world. Smith's death was an inevitable consequence of pre-existing circumstances.

But, relative to the information that you had, it was also quite random.

3.1 Determinism vs. randomness (continued)

Supposing that the world is *indeterministic,* would there for that reason be "random" events in it? The answer is "yes." But the reasons are not what one would think; and the amount of randomness in such a world is not necessarily as great as one would expect it to be.

For argument's sake, suppose we know that, given the current state of the universe, it's causally completely undetermined whether, in five seconds, object O will move to the right or to the left. Each possible outcome is as likely as the other (they're "equiprobable").There's no way, of course, to predict which way it will go. We could make that prediction only to that the extent that we had reason to think one of the outcomes the more likely one—which we obviously don't. But when, at the moment of true, it goes to the left, was that "random"? We couldn't predict it; but it isn't exactly random. On the contrary, it's consistent with our expectations of it.

To see why, let's talk about Smith again, but let's change the story. This time, he isn't killed by a ceiling lamp. The vindictive criminal kills him. Before Smith died, you knew that, in all likelihood, he would be killed *either* by that criminal *or* by his wife *or* by his heart-condition—but you didn't know which, since, given what you knew, they were all equiprobable. When the criminal gets him, that surely isn't random. Quite the opposite. Even though it could not itself be predicted with certainty to happen, it belonged to a class of events which, given the available evidence, was known to contain all and only the things that *would* kill him.

The same is true of O. In going to the left, it's doing something that, being uncaused and therefore not *itself* predictable, it belonged to a class known, or at least reasonably believed, to contain all and only the things that O might be doing in five seconds. So indeterminism doesn't entail randomness.

But here we must make a subtle distinction. *Relative* to the knowledge that it's undetermined whether O goes to the left as opposed to the right, or *vice versa,* it's going to the left is random. So O's going to the left is both random (relative to one body of knowledge) and non-random (relative to

another).

4.0 Why determinism is capable of being verified

On hearing me say that the world is at least approximately deterministic, students of mine often respond by saying either that determinism is unverifiable or that it's simply false. Depending how it's meant, this statements is either false or vacant. With some trivial exceptions, any universal statement (i.e., any statement of the form "all x's are y's"; e.g., "all birds have feathers") is incapable of being definitively verified. (The trivial exceptions concern cases where the generalization concerns only some very circumscribed class of objects; e.g., "all pencils on my desk are yellow," or where some statement that, in terms of its content, is not a universal generalization is turned into one through syntax-chopping; e.g., "all objects are such that, if they are identical with George W. Bush, they are mammals.") Since determinism is given by the statement "all events have complete causes," it is unverifiable, like "all metal expands when heated." On this basis, it is concluded that determinism couldn't possibly be known to be true, and shouldn't be affirmed, even tentatively.

But is it so clear that "all metal expands when heated" is unverifiable? So far as that is true, it's only in the trivial sense in which nothing other than analytic truths can be definitively verified. To be sure, "all metal expands when heated" couldn't possibly be verified on *strictly observation grounds.* After all, that statement concerns *all* metal, not just the metal you see. So no universal generalization, with the irrelevant exceptions noted a moment ago, can be verified on *purely* observational grounds; some element of inference is involved. *Anything* that can be observed to occur is, for that very reason, not general. And that is why, despite its syntactic similarity to "all metal expands when heated," "all pencils on my desk are yellow" is not *substantively* (as opposed to just syntactically) a universal generalization: one can *observe* that it is true. You can observe this or that metal bar expanding; but you can't observe the "general fact," if there even is such a thing, that heated metal always expands. Also, if by "verification" you mean "observation-based verification," then "all metal expands when heated" and "every event has a cause" are incapable of verification, but only because "verification" has been so defined as to render that statement an empty truism.

But if one allows at least some inference-rules to be permissible sources of verification, then "all metal expands" is verifiable, depending on what those rules are. And if those rules are anything like those that laypersons and scientists permits themselves, then "all metal expands when heated" *is* verifiable. So it is only in some trivial sense that "all metal expands when heated" is unverifiable.

4.1 Why determinism is capable of being verified (continued)

Some say that "all metal expands when heated" is unverifiable in the sense that, no matter how many metal objects you've seen expand when heated, the next one "could" fail to expand. But what does "could" mean in this context? There is a sense in which I "could" win the lottery, or even sprout wings, even though the latter and probably also the former are prohibited by natural law. (In fact, the former is definitely prohibited by natural law since, given who I am, it isn't psychologically possible for me to buy lottery tickets. I find the very concept of a lottery appalling.) To say that I "could" sprout wings is not to say that it's permitted by natural law, let alone that it's possible in some more robust sense. It is only to say that, given what I know, it isn't ruled out. But I would contend that I do know that I won't sprout wings and, therefore, that what I know does rule it out. And you too, outside of a philosophy class, would say that you know that you won't sprout wings.

David Hume famously argued that inductive inference isn't rational. His argument is not easy to refute. (That said, we saw in Chapter 12 how to refute it.) But nobody accepts Hume's conclusion; and while it's one thing to say that we don't know how to justify a belief, it's quite another to say that it has no justification. Surely my belief that I won't sprout wings in the next five minutes is not in the same category as a person's belief that tomorrow it will start raining gumdrops. And it's a veritable truism that the difference is that the former is justified, whereas the latter is not.

4.2 An oft-heard argument as to why determinism is incapable of being verified

Here is an argument that several students, independently of one another, have presented to me, and that is also found in the work of C.S. Peirce (1839–1914):

> Let cell X be a "perfect" clone of cell Y, and suppose that X and Y are kept under exactly similar conditions. (So, although they are kept in different Petri dishes, they are subjected to precisely the same forces.) X and Y will behave differently from each other. Therefore, the same initial conditions fail to lead to the same non-initial conditions. According to determinism, this isn't possible. (If two qualitatively identical circumstances are succeeded by qualitatively different circumstances, then neither circumstance completely determined what it would be succeeded by.) From this it follows that determinism is false.
>
> Given only that two cells are biologically exactly alike, it doesn't follow that they're exactly alike in all other respects. It doesn't follow, in particular, that they are exactly alike at the atomic and sub-atomic levels. And the chances of their being thus alike are basically zero. There is no real chance that the micro-particles composing one organism will be moving in precisely the same way as the micro-particles composing any other organism. These sub-biological facts can and will have dramatic effects on what happens at the biological level. All this argument establishes, then, is the irrelevant point that *different* initial conditions may lead to *different* non-initial conditions.

4.3 Another oft-heard argument as to why determinism is incapable of being verified

Many students of mine have put forth the following argument:

> Determinism says: "same initial conditions, same non-initial conditions." But the conditions obtaining at any given time are different in at least some respects from those obtaining at any other time. For, between any two points in time, there has surely been at least one change; some particle has moved or rotated or vibrated.[266]

So there is no way to verify determinism.

Causal connections are discovered by seeing how differences in initial conditions correlate with differences in non-initial conditions. You allow initial changes to change in *one* respect, but do your best to ensure that they're unchanged in every other respect. This will guarantee that these other factors will at least *tend* to remain constant, even though, inevitably, they'll never be perfectly replicated. If changes in that one factor reliably correlate with changes in the outcome, there's probably a causal connection; and there's probably no such connection if there's no such correlation.

You may not be able to establish a *perfect* correlation between your pushing the button and the elevator's coming. But that doesn't mean that you can't have good reasons for positing the existence of a strictly deterministic connection between the first and the second. For there to be such a connection is *not* for the elevator to come without fail every time the button is pushed. It is for the likelihood that the elevator will come, given that the button was pushed, to approach 100% to the extent that other factors are held constant; that is, it is for there to be no *limit* to how close to 100% that likelihood can be made to be by rooting out other factors that aren't constant.

Your experience can easily give you good evidence that *this* condition is satisfied. If you find that it's always possible to make it *more* likely than before that the elevator will come, given that the button was pushed—if, in other words, you find that, given enough tinkering about with the relevant circuitry and cables and so on, it's always possible to tighten this correlation —you *ipso facto* have evidence that the condition in question is satisfied and, therefore, that a strictly deterministic connection is at work.

This doesn't show they can be definitively established. But it *does* show that, insofar as they're incapable of being definitively established, it's only for the reason that *nothing* known through sense-perception can be definitively established. Your sense-perceptions make it probable that you have a television, but they don't make it certain. They make it probable that you're not a disembodied spirit, but they don't make it certain. They don't make anything about the external world certain. A fortiori they don't make the existence of any causal connections certain. But such connections are no less capable of being definitively established than anything else about the

external world. Whatever shortcomings induction may have as a form of inference, they're no more of a problem for attempts to establish causal connections than they are for attempts to establish that you have hands.

We can't have perfect evidence of a perfect correlation.[267] But that doesn't mean we can only have evidence for imperfect correlations. It's possible to have imperfect evidence for perfect correlations. I've examined a million mammals, and they all had backbones. Here we have evidence of a perfect correlation. It's imperfect evidence, but that's the only kind that's available. More precisely, it's imperfect evidence *if* by "perfect evidence" we mean definitive *observational* evidence—as opposed to observational evidence *plus* reasonable inference-rules. But if this is what we mean by "perfect" evidence, then it's a sterile triviality that we can't have perfect evidence for the existence of transperceptual correlations. (By "transperceptual," I mean "happening on the other side of the veil of perception—out there in the world, as opposed to in our minds.")

We have similarly "imperfect" evidence for deterministic laws. And such laws are no *more* incapable of being definitively established, and no less capable of being established with reasonable certainty, than biological concomitances of the kind just described.

Chapter 15

The Compatibilist Conception of Freedom

1.0 Determinism in relation to the debate between materialism and dualism

A set of events is deterministic if it obeys laws that, given the system's state at any given point in time, fix the system's state at all later points in time. The physical world is a deterministic system. (In any case, it's very nearly so. And in this context, we'll assume that it is. Later, we'll drop the assumption.) This means that the world itself is a deterministic system *to the extent* that it's physical.

There are both physical and mental entities. Anything that involves thought or feeling is mental; anything that does not (but is spatiotemporal; i.e., is in space or time) is physical. So beliefs, intentions, and regrets are mental; and rocks, trees, and explosions are physical.

It is an open and delicate question how the mental and physical are related. (See the Appendix to Chapter 11 for a discussion of this matter.) Some hold that mental entities are *identical* with certain physical entities (presumably neural events or structures). This position is known as ***materialism.*** Others deny this, saying that the mental is non-physical. This position is known as ***dualism.***

If materialism is right, then the mental world is just part of the physical world. So if E is some experience that you either have had, are having, or will have, then given the way the world was 10 million years ago, the state of the universe 10 million years ago made E's occurrence inevitable. Whatever you decide to do today is an inevitable consequence of the way the world was a million years ago.

Supposing that the physical world is deterministic, and supposing also that materialism is right, it's hard to see how there could be free will. If the state of the universe a million years ago completely settled what choices I would make today, then presumably I couldn't make any choices other than those that I *do* make. And it would seem that, under those circumstances, I don't have anything resembling "freedom of choice." In a moment, we will find that some case can be made that one may have *a* kind of freedom in a deterministic world. But let's hold off on that for a moment and talk right now about the relationship between *in*determinism and freedom.

2.0 Free will in an indeterministic world?

Because of the apparent irreconcilability of materialism with the presumption that we are free, some believers in free will—for example, Jean Paul Sartre (1905–1980) and Maurice Merleau-Ponty (1908–1961)—accept dualism, while others, e.g. Peter van Inwagen (1942–), accept materialism but reject determinism. Let's consider both options, starting with the first.

First of all, supposing that there is indeterminism in the world, it doesn't necessarily mean that our mental processes aren't determined. If the atomic processes making up some rock happen to be indeterministic, that doesn't

mean that your brain or mind aren't deterministic. So the indeterminism must be in the right place.

What is that place? Where must the indeterminism be if our choices are to be made freely? The indeterminism must relate to the processes involved in making choices. For, if the mechanisms of which our choices are results are deterministic, then our decisions are as determined as anything else. Thus, a free person's choices are undetermined.

But a choice that isn't predetermined by anything isn't predetermined by any facts about one's personality, and a choice of the latter kind is not one that we would have any inclination to regard as "free." A short story may illustrate my meaning. Smith is walking along. All of a sudden, he decides to become a pilot. This decision has no basis in his personality. He never liked flying. He already has a career that he enjoys and he depends on. So it would be inconvenient for him to become a pilot, and he would derive no pleasure from it. He knows all of this. Nonetheless, he makes this choice. It just comes out of the blue.

Was this choice made *freely?* It doesn't seem like it. It actually seems rather *compulsive*—like something that just overcame Smith. Suppose that "choices" similar to the one just described were a frequent occurrence in Smith's psychological life. Would we say that Smith was "free"? No. We'd describe him as having a kind of disorder. We'd describe the fact that he was visited by these strange compulsions as indicative of a lack of psychological integrity on his part; and see his submission to these forces as showing a *lack* of freedom on his part. Smith is a puppet dancing at the end of strings pulled by psychopathology.

This suggests that, in order to be made "freely," a choice has to be rooted in facts about one's personality, and must express one's values, desires, beliefs, and previous intentions. A choice that lacks these roots seems to involve a disintegration of the person's psyche and, paradoxical though this may seem, a consequent diminution of the subject's freedom. So even if there is indeterminism, it doesn't help people be free; indeterminism actually seems to hurt our chances of being free.

But to the extent that a choice *is* appropriately rooted in a person's personality, that choice is *not* the result of indeterminism but, on the contrary, of a deterministic structure. And it's hard to see how, under those

circumstances, we could be free. So we're damned if we do and damned if we don't!

3.0 Free will in a dualistic world?

Now let's consider the view that, although the physical world is deterministic, the mental is not physical. First of all, even if the mental is not physical, there is considerable evidence that the mental sphere is deterministic. The decisions that one makes are clearly a product of one's pre-existing mental condition.[268]

We must also take into account that, at least according to some viewpoints, the choices we make have extremely deep roots in our unconscious mental lives. Analysis tends to suggest that decisions that seem, at the level of consciousness, to be "random" are, in fact, strictly determined by unconscious forces. In any case, this is one point of view, and it is one for which there is some corroboration.

But suppose, if only for argument's sake, that your choices were not fixed by the state of the universe prior to their occurrence. For the reasons given earlier, that wouldn't make those choices be free. On the contrary, it would make them more like fits or seizures than like considered, freely made judgments.

A story similar to the one just told may help make this clear. You are a student at State University. You have an A-average. All of a sudden, you choose to drop out of college and become a drug dealer. Your choice was not caused to occur by anything; the state of your personality prior to the choice had nothing to do with your making it; neither did the state of the universe. It just happened. So, in the middle of a test, you throw down your pen, bolt out of the room, and get to work as a drug dealer.

Does this sound like a freely made choice? No. It sounds more like a compulsion. It's as though you were "taken over" by another person and lost control of your body. Again, we see that indeterminism doesn't help freedom and actually undermines it.

4.0 The Compatibilist Conception of Freedom

The problem is that, as we saw, determinism *also* seems to threaten freedom. Obviously you have no control over the past. If determinism is right, then what happened in the past completely determines exactly what's happening in the present and, in particular, completely fixes what choices you're making now. So something over which you have no control is completely responsible for the choices that you're making. It's hard to see how, under these circumstances, you could be said to have free will.

Nonetheless, some hold that there *can* be free will in a deterministic world. Such people are known as ***compatibilists.*** Those who deny it are ***incompatibilists.***

4.1 The crude form of compatibilism

There are ultimately two different forms of compatibilism. One of these is very straightforward; the other is quite sophisticated.

According to the straightforward version, which we'll discuss in this section, you are free to the extent that you can do what you want, and you are unfree to the extent that you cannot. If you want to leave your room and you can, you are free. (In any case, you're free *in that respect.* You may be unfree in other respects; e.g., once you leave the room, you may want to leave the house you're in but not be able to.) According to the other version of compatibilism, which we'll discuss in Section 4.2, being able to do what you want, though necessary for freedom, isn't sufficient for it. It is necessary also that your choices be adequately rooted in who you are as a person.[269]

4.1.1 G.E. Moore's compatibilism

According to G.E. Moore (1873–1958),[270] an act is performed freely if the agent *would* have done something else *if* he had chosen to do so. An act is free if the agent "could have done otherwise."

The idea is pretty straightforward. Although I'm considering striking Harry, I choose not to. But, immediately upon making that choice, an extremely strong person grabs my arm and forces me to hit Harry. My act was clearly unfree. But why? Because, says Moore, even though I chose to do otherwise, I ended up striking Harry. Now suppose that I *voluntarily* strike

Harry. I do it, not because somebody forces me to, but I don't like Harry and want him to suffer. In this case, my act *is* free. But why? Because, says Moore, if I hadn't chosen to strike Harry, I wouldn't have done so. Thus, my striking Harry is voluntary, Moore says, iff I wouldn't have done so had I chosen not to do so.

Moore's position is a form of compatibilism. Free acts aren't those that are uncaused; they're those that are caused in a certain way—they're those that are caused by decisions on the part of the people (or animals) who commit them.

4.1.2 Chisholm's objection to Moore's analysis

In *Human Freedom and the Self,* Roderick Chisholm (1916–1999) makes a compelling objection to this analysis. Let's suppose that, if I had chosen not to strike Harry, I would indeed have refrained from doing so. Unless I was at liberty to *not* make[271] that choice, it's hard to see how my doing so was free. If every choice I make is inevitable, then it's irrelevant that, had I made different choices, I would have acted differently. For I *couldn't* have made those choices under those circumstances, and I therefore *couldn't* have acted in those ways. And if I couldn't have acted in those ways, then I couldn't have done anything other than what I did, making it hard to see how, in acting as I did, I was acting freely.

4.1.3 A Lockean counter-objection

Chisholm's point is a good one, but it isn't decisive. John Locke (1632–1704), who held a view similar to Moore's, makes it clear that points like Chisholm's don't necessarily carry the day.[272] In Locke's view, freedom is a property of *actions.* It makes no sense to speak of anything other than actions as being free. Only things that one can *intend* to do are done freely. I can *intend* to do 20 push-ups, and that's why I can do it freely. But I can't *intend* to make a certain choice. What would it be to *intend* to make a certain choice (e.g., to intend to do 20 push-ups)? If I really *intend* to choose to do 20 push-ups, then, for that very reason, I *have* chosen to do 20 push-ups. Second-order intentions (intentions to intend) collapse into first-order

intentions. Thus, so far as it isn't a pleonasm, the expression "intentional choice" is meaningless.

Locke's argument is a good one. But in a while, we'll see that, according to one, quite defensible form of compatibilism, choices *can* be made freely—that, in fact, freedom is *primarily* a property of choices and only secondarily of acts.

4.1.4 Frankfurt's refutation of Moore: why "x acted freely" ≠ "x could have done otherwise"

Harry Frankfurt (1929–) decisively demonstrated the falsity of Moore's position by identifying scenarios in which somebody performs some act freely even though that person could *not* have abstained from performing that act.[273] Here is such a scenario:

> (SM[274]) Mr. Smith intends to kill Mr. Jones. Like Mr. Smith, Mrs. Green also wants Mr. Jones dead. But Smith doesn't want to go to jail for murder. Without Smith's knowing it, she installs a microchip in his cranium. (We needn't worry about how exactly she does this or how exactly the microchip works.) By means of this microchip, she can control Smith's actions—she can make him do whatever she wants. She knows that Smith is intent on killing Jones. But if Smith should get cold feet, Mrs. Green will use the microchip to force Smith to go through with it. Smith doesn't know about the microchip; nor does he even know of Green's existence. Smith kills Jones—he doesn't chicken out at the last minute. And he does so freely (at least insofar as anyone does anything freely—he is not in the grips of any delusions or compulsions, etc.).

Thus, Smith freely kills Jones, even though Smith could not have done otherwise. For had he decided at the last minute *not* to kill Jones, Green would have forced him to.

Hence, contrary to what Moore alleges, "x performs act A freely" is *not* equivalent with "x does A and, had x chosen *not* to do A, x would not have done A."

4.1.5 The irrelevance of Frankfurt's analysis

But even though, interpreted narrowly, Moore's analysis is false, it's really just a misguided way of stating the principle, which Frankfurt's argument does nothing to undermine, that a free act is one that is choice-driven. Moore was led to equate "x acted freely" with "x could have done otherwise" *not* by a misconception concerning the nature of freedom, but by his acceptance of a certain analysis of causality—the so-called "***counterfactual analysis of causality***" (CFA). According to CFA, x caused y to occur just in case if x hadn't happened, y wouldn't have happened. "Smith's pushing the button caused the elevator to come" means the same thing as "if Smith hadn't pushed the button, the elevator wouldn't have come."

CFA is wrong. This is because there is causal redundancy in the world. I push the elevator button, causing the elevator to come; but if I hadn't pushed it, one of the other 20 people waiting for it would have done so, and the elevator would have come.[275]

But a close cousin of CFA is correct, namely:

(CFA*) x caused y to occur just in case, if x hadn't happened, then ***ceteris paribus*** y wouldn't have happened.

Similarly, even though, for the reasons just given,

(PAP[276]) "x did A freely if and only if x did A and, further, had x chosen not to do A, x wouldn't have done y"

is false, nonetheless

(PAP*) "x did A freely if and only if x did A and, further, had x chosen not to do A, then *ceteris paribus* x wouldn't have done A"

is correct.

It's well known that, if they are to come out correct, counterfactuals typically have to be hedged with *ceteris paribus* clauses. Even though:

(JFK) "if JFK hadn't been assassinated, he would have been re-elected"

is a plausible counterfactual, it isn't necessarily true. What if, sometime after

Nov. 22, 1963, but before the '64 elections, JFK died of natural causes or was disgraced by some scandal? In that case, he would not have been re-elected. So the right counterfactual is:

(JFK*) "if JFK hadn't been assassinated, then *ceteris paribus* he would have been re-elected."

In other words, if JFK hadn't been assassinated, but *otherwise e* things were the same, so that JFK was healthy and popular, etc., *then* he would have been re-elected.

Notice that JFK* is really a causal claim, since it's just another way of saying:

(JFK#) "JFK's being assassinated caused an otherwise re-election-bound JFK to *not* be re-elected."

As a rule, counterfactuals are causal claims.[277] Thus, "Smith killed Jones freely" is true just in case (i) Smith killed Jones and (ii) if Smith had chosen *not* to kill Jones, then *ceteris paribus* Smith would have not killed Jones.

Basically, Smith killed Jones freely if his doing so expressed his decision to do so, as opposed to some other force. In general:

(FA[278]) an act is free if it expresses one's decision to act that way.

This is clearly the right analysis of *a* kind of freedom.

4.1.6 Freedom not a property of choices

If FA is right, it is acts, not *choices,* that are free. This is what Locke held. Locke was the first advocate of FA, and he was an exceptionally effective one. One chooses to act, Locke said; one doesn't choose to choose. It makes no sense to suppose that one chooses to choose. For if one always chose what to choose, Locke points out, one would have to make infinitely many choices in order to make a single one, and one therefore couldn't make any choices at all. In order to choose to open the door, one would have to choose to choose to open the door; and in order to choose to choose to open the door, one

would have to choose to choose to choose to open the door; and so on.

So freedom is freedom to act, not freedom to choose. As Locke puts it, freedom is a relationship between choice and act, and not between subject and choice.

It follows that, so far as one's freedom is concerned, it is irrelevant whether one's choices are caused or not. Since freedom is a relationship between choices and acts, not between choices and their own antecedents, it doesn't matter whether choices are caused. Nor, therefore, does it matter *how* they're caused.

In fact, if FA is right, determinism is not only compatible with freedom, but is a prerequisite for it. Wanting to make the elevator come to your floor, you intend to push the button. If a degenerative neurological condition has undermined the integrity of the mechanisms linking your brain to your hand, thereby making them indeterministic, your intention to move your hand may be more likely to fail than to succeed. And, supposing that you do manage to push the button, if the mechanisms linking the button to the elevator are indeterministic, your pushing the button may be more likely than not to fail to get the elevator to come. To the extent that the mechanisms involved in the implementation of our choices are indeterministic, we are in the grim position of somebody who has Parkinson's. The spirit is willing, but the flesh can't be counted on to obey, and neither can the relevant mechanisms in the external world.

4.2 Frankfurt's analysis of freedom[279]

According to Locke, you are free to the extent that you can do what you want to do, and you're unfree to the extent that you can't. This is obviously *a* kind of freedom—a very important kind.

But it isn't the only kind. And, contrary to what Locke says, there is *a* sense in which choices can be free or unfree. Harry Frankfurt made this clear.[280] A story will help us begin to understand Frankfurt's insightful analysis of freedom.

Jones is a great writer who is addicted to heroin. Because his addiction is only moderate, he is able to function perfectly well even when he is heroin-

deprived. (When heroin-deprived, he doesn't feel good, but he is in no way incapacitated.) He knows that, if he were to shoot up right now, the subsequent lethargy would make it impossible for him to write a single word for the rest of the day. He also knows that, if he doesn't write 10 pages today, he will lose his book contract. Jones does not want to lose the contract. Plus, he wants to grow as a writer, which he won't do, at least not today, if he shoots up. He thus has a desire to *not* take the heroin—even though he simultaneously has a desire *to* take the heroin.

Jones' desire to take heroin is a first-order desire. Jones' desire *not* to act on this desire is a ***second-order desire.*** This is because Jones' desire not to take heroin is a desire as to what to do with another desire; it is a desire *not* to cave in to a desire to take heroin.

Suppose that Jones caves in and he takes the heroin. In that case, Jones is a slave to his desires. His will has been compromised. He is not acting freely.

But *now* suppose that Jones does *not* cave in. He does the right thing: instead of taking the heroin, he endures the discomfort of a heroin-free day and manages to get his work done. In *this* context, Jones is not a slave to his desires. His will has not been bent. His conduct expresses a "free will."

In light of considerations such as these, Frankfurt says that somebody has a free will to the extent that his conduct aligns with his *second*-order desires.

But here a question arises. Why are second-order desires so special? Why is it that freedom consists in doing the bidding of one's *second*-order desires —as opposed to that of one's third-order, or fourth-order, or first-order desires? What, in this context, is so special about the number two?

When the heroin-addicted writer decides *not* to take heroin, what is driving him is a rational appraisal of his situation. But when the heroin-addicted writer gives in to his addiction, it is *not* his rationality (or his morality) that is driving him.

Before we can close this argument, we must address Frankfurt's discussion of the concept of a "person." By a "person," Frankfurt does not mean a member of the species *homo sapiens.* He means anything that has the same basic psychological architecture that we have. What is so special about that structure? First of all, we are self-conscious. Second, we are rational.

Third, we are rational not only about how to attain our objectives, *but also about what those objectives should be.* What distinguishes persons from non-persons is that the former, but not the latter, subject their own desires (and, more generally, their own mental states) to critical scrutiny.

Non-persons are often very clever when it comes to figuring out how to attain their objectives; for example, beavers show a lot of intelligence when it comes to building dams. But they are not smart about whether they should be building dams in the first place. The question doesn't arise. They have the desire, and they act on it (or they die trying). In general, persons do, whereas non-persons do not, subject their own desires to rational scrutiny. Non-persons do not ask themselves: "What would the consequences of acting on this desire be? And *should* I act on this desire?"

We persons, on the other hand, *do* subject our objectives to rational assessment. And this is obviously made possible by the fact that we are self-conscious. If I am overtaken by a desire to build a dam, I will ask myself: "how would my acting on this desire help my overall interests?"

We can now close the argument. Whenever someone is driven by a rational assessment of their own desires, they are driven by a desire about a desire. Thus, one is rational only to the extent that it is one's second-order desires that drive one's conduct. There is a clear sense in which rational conduct is free and in which irrational conduct is not.

There's more to being free than being able to do what you what. Your desires must reflect who you are. If they don't, those desires *belong* to you but aren't *of* you. And in that case, you are alienated from your own mental states, in much the way that a person with Parkinson's is alienated from his body.

But, in a way, the alienation is even worse. Involuntary behavior is unintentional behavior. Intentional behavior is voluntary behavior. Voluntary behavior is desire-driven behavior. (If you don't *want* to raise your arm, but it goes up anyway, you didn't intend for it to happen.) Epileptic fits are not *actions.* Unintended behaviors are not *action.* Unlike actions, they're *undergone,* not *done.* So estrangement from one's own actions amounts to estrangement from all of one's actions; and it's hard to imagine a condition less free than that.

4.2.1 Frankfurt's emendation of this analysis

Another story will help expose a problem with the analysis of freedom just described, and it will also clarify the raison d'être for Frankfurt's soon to be discussed emendations of it.

Smith's first-order desires are seriously out of alignment with his second. He values writing poetry; he's good at it; and he enjoys it about as much as it's humanly possibly to enjoy it. But he has a voracious craving for drugs. His craving for drugs, especially cocaine, is unusually strong; and his desire to write poetry, though intense, is not as intense as his desire to do cocaine.

Smith's desire to do cocaine is out of alignment with many of his values (second-order desires). Were he to give into this desire, he couldn't write poetry; nor could he maintain friendships or play sports or, with only a few exceptions, do any of the other things he values. For this reason, he doesn't give in. But given how intense his craving is, every day is a struggle for Smith. If, as Frankfurt suggests, one is free to the extent that one acts on one's second-order desires, Smith is as free as a person can be. But given how hard life is for Smith, it's counterintuitive to suppose that he's maximally free. To be sure he has *a kind of* freedom, and he has it in abundance. But there's some other sort of freedom that he doesn't have.

Frankfurt is aware of this problem, and he revises his view to deal with it. According to his revised view, if an act is to be free, it is *necessary* that it be driven by a second-order desire, but it isn't *sufficient.* Another condition must be met: one's first-order desires must be consistent with the second-order desires that are driving that act.

Here's the idea. Person X has no self-control. He binges on drugs and food, and he hates himself for doing this; but he can't stop—he's too weak and his urges are too strong. His life has no meaning and he knows it; but he will not, or cannot, do anything about it. Unlike X, Person Y has a huge amount of self-control and never caves into temptation. But everything is a struggle, since he's always fighting temptation, like a drowning man trying to stay afloat. (Y is like Smith.) Person Z cannot be categorized with X or with Y. Z always does what he values, which is composing music. And what he values is what he *wants.* He wants what he wants to want. Because he's doing what he wants, he has the freedom that X has that Y lacks. Because he's

doing what he wants to want, he has the kind of freedom that Y has that X lacks. And since, of the three, he's the only one who has both kinds of freedom, he's obviously the most free of them all.

Thus, one is free insofar as two conditions are met: (i) one acts on one's second-order desires, as opposed to one's first-order desires; and (ii) one's second-order desires are in alignment with one's first-order desires.

To see the merits, and also the shortcomings, of Frankfurt's brilliant (and, I think, largely correct) analysis, we must take a moment to say how that analysis relates to compatibilism.

4.2.2 "x is free"= "x does what x wants to *provided* that what she wants to do is adequately rooted in her personality structure"

We've considered two analyses of freedom: (i) "x performed act A freely" = "it is because x chose to A that he did it," and (ii) "performed A freely'' = "A reflects who x really is." These two views can be combined into the view that (iii) "x performed act A freely" = "it is because x chose to A that he did it, *provided* that this choice of x's reflects who x really is."

This analysis, which is basically Frankfurt's, embodies a profound insight into the nature of human freedom. But, despite everything said thus far, a case can be made that at the end of the day *all* freedom is freedom to do what one wants. For reasons that we'll soon discuss, I myself don't accept this idea. But it isn't by any means absurd. Here's why:

> (FD[281]) Smith has given himself over to drug addiction. He spends days doing drugs. (He has unlimited supplies of money and never has trouble getting them.) Smith has many talents—he's a great water-skier, a great short-story writer, etc., but, because of his addition, these gifts of his lie fallow.
>
> Green, on the other hand, hasn't given himself over to addiction. He spends his days doing things that augment who he is. Green can't indulge every craving that he has. Unlike Smith, he doesn't have the money to do so.

You (JMK/Frankfort) would say that Green is freer than Smith, and your reason would be that, even though he can't gratify as many desires as Smith can, the desires that Green does gratify have the right roots, whereas those that Smith gratifies do not.

I grant that Smith's condition is worse than Green's, and also that Smith is less free than Green. But, contrary to what you say, that *is* because, when it comes to desire gratification, Green's batting average is better than Smith's. There are a lot of aspirations that Smith has that will be thwarted as long as he's imprisoned by addiction. Smith wants to do cocaine *more* than he wants to be a writer, but it doesn't follow that he *no longer* wants to be a writer. He could very well want both. [282] Let's suppose he does. In that case, a very strong desire of his is being frustrated. It's true that *he's the one who's frustrating it,* but that doesn't mean it isn't frustrated.

Smith will inevitably have many other desires that his cocaine addiction dooms to frustration. In his pre-addiction days, we may suppose, Smith enjoyed many things that he no longer enjoys. We can assume that he *desired* to do each of these things, and we can also assume that he *still* desires them, even though he obviously doesn't desire them as strongly as he desires cocaine. Many desires are rooted in personality structures. A desire to play the piano or read Nietzsche doesn't just overtake one; it's rooted in enduring structural facts about one's personality. And as long as the personality structure is there, one's ability to express it in the form of action is, practically by definition, a form of frustration. Smith's personality structure may still support those desires. (And almost certainly will. Addiction doesn't turn people into complete vegetables, at least not right away, and usually not ever.) It could be that Smith *thinks* it would be greater. But it's not so clear that it would in fact be greater. The phenomenological expression of that frustration would be more intense. that doesn't necessarily mean that the desire in question is itself more intense. But that doesn't necessarily mean that the desire in question is itself more intense. Desires, even short-lived ones, are structures; they aren't *feelings.* And the feelings to which a desire gives rise don't necessarily make it clear how intense that desire really is.

FD is close to the truth, but it falls just short. Another story shows why. Aaronson is a coward who has no self-confidence. So even though he wants to write novels and is extremely talented at it, he doesn't do so. It isn't safe enough. It's too bohemian for his stodgy, blinkered parents and for his equally stodgy friends (who he doesn't even like—he just pretends to like them, since he's supposed to). To make a living, Aaronson sells insurance, which he hates. His life is a waking death.

In living as others want him to, Smith is surrendering his freedom, *and* he is also dooming many desires of his to frustration. But the two, though related, aren't identical. Aaronson is suppressing an important side of himself. The personality structures that he is suppressing *give rise* to many ardent desires, but they aren't *identical* with such desires. His losing his freedom consists in his suppressing these structures. The desire frustration that follows is an *effect* of this loss of freedom and isn't *identical* with it.

So Aaronson's loss of freedom isn't identical with his being unable to do what he wants; his being unable to do what he wants is a by-product of his loss of freedom. To be sure, his inability to do what he wants is *itself* a loss of freedom. But, important though this second kind of freedom is, it's distinct from the first kind and a derivative of it. So being free consists primarily in being oneself, and secondarily in being able to do what one wants. And this is obviously close to what Frankfurt is saying, if it doesn't coincide with it.

4.3 Sartre and Merleau-Ponty on freedom

Jean-Paul Sartre (1905–1980) and Maurice Merleau-Ponty (1908–1961) both give what appears to be the same argument, to wit:

> (SMP_1[283]) Smith is in a locked room, and Jones is in an unlocked room. It is only to the extent that Smith wants to leave the room that his being unable to do so is an abridgment on his freedom. So, to the extent that Smith's being locked inside the room limits his freedom, it's because, to that extent, the world isn't complying with his will.
>
> The same points *mutatis mutandis* hold of Jones. It's only to the extent that Jones *does* want to leave the room he's in that his being able to do so constitutes a freedom. If he doesn't want to leave, his

being able to do so isn't a freedom, or is only a hollow one. To the extent Jones *does* want to leave the room, his being able to do so *does* entail that he's free (in that respect). But it isn't because his *will* is free that Jones is (in this respect) free. It's because the world meets his will halfway.

The difference between Jones and Smith isn't that the one has a freer will than the other. It's that the world meets the one person's will halfway but concedes nothing to the other person's will. They both have free wills. To have a free will is to be free to *decide* what you want. It is not to be free to *have* what you want.[284]

Supposing for argument's sake that SMP_1 is correct, what follows isn't that Smith's will is free, but only that his will is not made *un*free by the fact that he cannot leave the room. Locke would say that neither Smith nor Jones has a will that is at all free, the reason being, Locke would say that freedom is a property, not of the will per se, but of the relationship between will and act.

But even if, contrary to what Locke says, freedom *is* a property of the will, there is another reason to reject SMP_1. An extension of the story just told will help us identify this reason. Brown (some third person) has surgically implanted a microchip in Smith's brain. Brown controls Smith's desires and subsequent decisions by means of that microchip. So if Brown wants Smith to want to play tennis, then Brown can make Smith have that desire by means of the microchip. The locked room that Smith is now occupying is packed with beautiful artwork. Because he is an artist, Smith would ordinarily want to spend hours studying these masterpieces, and he thus wouldn't want to leave the room any time soon. But, using the microchip, Brown makes Smith desire to leave the room immediately—that is, Brown makes Smith's will be to leave the room immediately. But, at this instant, Brown unlocks the room, giving Smith the ability to leave.

Here we surely have a case of an *un*free will if ever there was one. So while it is true that, when he didn't want to leave the room, Smith's will was not made unfree by the fact that the door is locked, it doesn't follow that Smith's will was free.

Although neither Sartre nor Merleau-Ponty consider this exact point, they both put forth an argument that constitutes a kind of response to this, viz.:

> (SMP_2) Once again, suppose that Smith is in a locked room and wants leave it. There is obviously *a* sense in which Smith is unfree. If Smith had no desire to leave the room, he would not be unfree (in this particular respect).
>
> You are not able to turn into a tadpole. But this doesn't constitute an encroachment on your freedom, since you have no desire to be a tadpole. (If you did, it would.) The same is true of Smith. It is only *relative* to certain choices that he has made that Smith's freedom is limited. If he had no desire to leave the room, then his freedom would not be limited by his inability to leave the room.
>
> In general, it is only relative to the choices that we make that our freedom is limited. So there can be a *lack* of freedom only where choices have *already* been made. Since, consequently, the making of choices must occur *before* there can be any lack of freedom, it follows that our choices themselves are made freely. (Cf. Merleau-Ponty, p. 469: "[W]hat are called obstacles to freedom are in reality deployed by it. An unclimbable rock face, a large or small slanting rock, are things which have no meaning for anyone who is not intending to surmount them..." See also Sartre, p. 463.)

Given a Lockean conception of freedom ("to be free is to be able to do what you want to do—nothing more") SMP_2 goes through. So SMPT is conditionally correct. But on the basis of it, Sartre and Merleau-Ponty hold a view that is unconditionally incorrect. They hold that, given the truth of (SMP_2), it follows that choices are entirely *uncaused*.[285]

This is a major non-sequitur. It may be true that one cannot meaningfully speak of an absence of freedom except where choices have already been made. But it doesn't follow that those choices were uncaused or, consequently, that those choices were not predetermined. We've already seen why.

4.4 Ego-dystonic vs. ego-syntonic causes

To be free, we have argued, one must identify with one's desires—gratifying them isn't enough. Psychologists refer to mental contents that one identifies with as ego-*syntonic* and to those that one doesn't identify with as ego-*dystonic.* An understanding of this distinction may clarify why determinism is necessary for freedom and why indeterminism threatens it.

Some people who suffer from psychological disturbances *identify* with their symptoms. They don't see their symptoms as symptoms. For example, the acute schizophrenic who sees goblins dancing on his bed thinks that he's seeing things as they are. To his mind, it's other people who are wrong: they can't see what's there; he can. The same holds of the paranoiac who thinks her phone conversations are being monitored by the President.

These are ego-syntonic disorders. The viewpoint embodied in the patient's symptoms coincides with the viewpoint of the patient himself. The latter doesn't regard the former as an "alien intruder."

But not all psychological disturbances are like this. The classic example is obsessive-compulsive disorder (OCD).[286] Those afflicted with this disorder obsess over matters hardly worthy of their attention; they're compelled to perform absurd acts (e.g., they're compelled to snap their fingers if they have a thought about the number seven); they're also compelled to perform normal acts an abnormal number of times or for an abnormally long period (e.g., they're compelled to brush their teeth 20 times a day or to shower for hours on end). But sufferers of OCD *know* that their obsessions and compulsions are irrational. They succumb to the compulsions *not* because they actually believe they believe there to be any objective reason to do so, but because, for reasons unknown to them, they suffer intolerable anxiety if they don't. And they know that their obsessions are not legitimate; they know that those obsessions aren't comparable to Einstein's obsession with identifying framework-invariant physical laws or Beethoven's obsessive dedication to perfecting the Hammerklavier Sonata. They obsess despite themselves. They don't think their obsessions have any legitimate basis. They just can't stop them.

Thus, those afflicted with OCD don't identify with their symptoms.

Unlike schizophrenics, they see their symptoms *as* symptoms. Their symptoms belong to *their minds* but not to *them;* they are like alien-invaders that have taken root in their psyches.

Bearing these points in mind, consider the following scenario. You hate numbers. You are not detail-oriented. You want to spend all day writing novels. (You're a very talented novelist.) You know all of this. But one day you are overcome with an intense desire to become an accountant. This desire came out of nowhere. It has no roots in your personality.

You obviously don't identify with this desire. This is because it has no roots in who you are. And the reason it has no such roots is that your existing values, desires, aspirations, and so on, had no hand in giving rise to it.

Were you to act on this desire, there is a sense in which you'd be acting freely, but there's a deeper sense in which you'd be acting unfreely. You'd be acting (un)freely in a sense similar to (though not quite coincident with, as we'll see) the sense in which somebody who hands his money over to a gunman is acting (un)freely.

Suppose that, while pointing a gun at Smith, Green says "your money or your life." In handing over his money, Smith is acting freely in the sense that he's doing what it is that, the circumstances being what they are, he wants to do; and he's acting unfreely in the sense that, the circumstances being what they are, what he wants to do is not what he wants to want to do.

Much like Smith, you're damned if you do and damned if you don't. Supposing that you knuckle under to your ego-dystonic desire to become an accountant, you are acting freely in the sense that you are doing what you want to do, and you are acting unfreely in the sense that you are not doing what you want to want to do.

That said, there is a sense in which you are less free than Smith. The circumstances being what they are, Smith must exercise his agency in ways that he'd rather not to have to exercise it. But this agency is quite intact.

But *your* agency is not intact. Your ego-dystonic desire to become an accountant is competing for control over the very mechanisms that convert ego-*syntonic* intentions of yours into action; and were you to give into this desire, you would to that extent be forfeiting control over those mechanisms.

By virtue of having that desire, you're in a position similar to that of the President of a country much of whose is controlled by war-lords who don't recognize his authority—warlords who, though they operate *in* his country, are not *of* it, and whose sphere of authority is only as big as the President's sphere of influence is small.

Smith's situation is like that of a President whose country is being attacked by a foreign nation but who has complete control over his own country. Even though Smith is in a situation in which he is forced to act in ways in which he wouldn't otherwise act, his autonomy is not on that account compromised; and there is thus an important sense in which Smith is freer than you are.

4.4.1 A point concerning mental illness

Earlier, I referenced Beethoven and Einstein. Interestingly, both suffered from OCD.[287] So do many other high achievers. What this shows is not that OCD *per se* conduces to high performance, but that the personality traits that, when pathologized, are expressed as OCD lend themselves to high achievement. Beethoven, Einstein, and other OCD-afflicted high achievers were capable of intense concentration, were highly self-critical and, in order to satisfy the draconian standards of perfection they imposed upon themselves, were punctilious in their prosecution of the projects they undertook. Someone with these performance-enhancing character traits is more likely than others to respond to anxiety and disturbances generally by becoming obsessive and compulsive. But *mere* obsessiveness and compulsiveness are not pathological.

Also, there isn't a well-defined line between pathological and non-pathological obsessive compulsiveness; and one's pathologies may be implicated in, and in some cases even enhance, one's efforts. So given that Beethoven had OCD, it's unlikely that it was to no degree involved in his composing the Hammerklavier Sonata, even though his composing that work was, on balance, non-pathological.

This exposes one of the main differences between physical and mental illness. Physical illnesses are to no degree adaptive (except in some purely accidental manner—one's having asthma might prevent one from joining the

army and then being injured on the battlefield[288]). But mental illnesses, specifically ego-dystonic ones, are sometimes adaptive. This is because ego-dystonic illnesses are what result when defensive mechanisms become too strong. Excessive insecurity, fearfulness, and anxiety are defense mechanisms that have gone too far and that, consequently, inhibit the patient in some respects. But those same overdeveloped mechanisms may enhance the patient in other respects. An OCD mathematician's ability to focus on a minute task for hours on end is conducive to real achievement. In that respect, it's non-pathological. And there's no real chance that the psychological mechanisms that make it possible for the mathematician to sustain such a high level of concentration for such a long time have a *purely* pathological basis. But it's a distinct possibility that, depending on the mathematician, the mechanisms that make that possible are considerably reinforced by mechanisms the very purpose of which, at least when they were first developed, was to narrow the individual's psyche, so as to cast intolerable conceits into the abyss of the unconscious.

It should be pointed out that mechanisms whose *original purpose* is neurotic can end up becoming non-neurotic virtues of character. For example, to deal with some specific problem (e.g., a sense of helplessness that is induced by an unstable home-life), a given person might over-develop some existing character trait (e.g., tendency to focus on details becomes a maniacal obsessions with them), which, after the initial disturbance, has been dealt with, assumes some non-pathological, adaptive form (e.g., a talent for neurosurgery).

Basically, there isn't always a clear distinction between mental illness and mental hyper-wellness. And the fact that the border is blurry isn't to be understood in terms of the trivial, ubiquitous fact that most borders are blurry (that, for example, there is no sharp line between short and tall, poor and rich, etc.). The blurring of the line, where mental illness is concerned, has to do with the fact such illness is, in some ways, at certain junctures, constitutive of certain kinds of mental health (or, what may be different, mental flourishing).

This, I think, is why "genius" and "madness" are proverbially thought to go together. I *don't* think that the two, in fact, go together. Madness is always performance-inhibiting, for reasons that we'll soon discuss. But the

eccentricities associated with ego-dystonic ailments are often conflated with madness, leading to the false view that insanity is associated in some way with genius. I must emphasize that ego-dystonic ailments are *not* a form of madness. (Not all psychological disturbances are forms of madness.) So far as the word "madness" has a clear meaning, it refers to a condition in which one *identifies* with one's pathologies—it refers to ego-syn-tonic ailments. A precondition for such conditions is a lowering of inhibitions that, in sane people, remain operative and that, when *hyper*-operative, constitute OCD and other ego-dystonic ailments. Goblins and ghouls are no part of a sane person's life except when, during sleep, there is a lowering of the inhibitions that, during wakefulness, rightly keep these creatures of childhood in the cellar of the unconscious. A person with OCD (who isn't also a person with an ego-syntonic disorder) is like the sane person, except that the OCD person's defenses are in overdrive. So, in such a person, it isn't just ghouls and goblins that get the ax: healthy beliefs and sentiments do as well. One's antipathy towards one's father, and one's suspicions concerning one's best friend become off-limits in the way that, in a less neurotic person, only goblins and dragons are.

In terms of its outward effects, this hyper-inhibitedness can mimic "madness" (which, so far as it means anything, refers to psychosis). The psychotic person refuses to touch the doorknob because he thinks it's Satan's horn. The obsessive-compulsive person refuses to touch it because, in his mind, it's connected, by way of emotionally charged associations *that he knows to have no objective basis,* with emotions that he doesn't countenance (e.g., aggressive urges towards a loved one), sexual urges that are "wrong" (e.g., homosexual, incestuous, or sadistic urges). But the outward observer only sees some eccentric who refuses to touch a doorknob. Hence the conflation of psychosis with ego-dystonic illness; and—since the latter is associated with (though it's not the most productive manifestation of) character traits that are associated with high achievement—hence the bogus stereotype of the "mad genius."

So, ego-*syn*tonic illnesses, unlike ego-dystonic ones, are not at all adaptive, the reason being that they result from a *weakening* of one's defenses to a profound, highly incapacitating disorganization of one's psyche.

So, in a way, those suffering from ego-dystonic illnesses are even further removed from those suffering from ego-syntonic illnesses than those who suffer from neither. In this connection, it is worth pointing out ego-dystonic maladies seem to coincide with what are sometimes called “neuroses” and that ego-dystonic ones coincide with “psychoses.”

Chapter 16

Personal and Objectual Identity

1.0 Identity-related expressions ambiguous

We often make statements of the form ‹x is identical with y›. Examples are “I’m the guy in the photo,” “water is H_2O,” “Superman is Clark Kent,” “that’s the eraser I was looking for,” and so on. Statements of this form are known as “identity statements.”

Identity statements don’t always, or even typically, use the terms “identity” or “identical with.” Usually, they just use some form of the verb “to be.” So “that’s the guy who took my money” is an identity statement, even though the terms “identity” and “identical with” are nowhere to be found in it.

1.1 Identity-related expressions ambiguous: predicative vs. non-predicative uses of such expressions

This brings us to a fact that confused many pre-20th-century thinkers. The verb “to be” is ambiguous. As just noted, it sometimes means “identical with” (“he is George” means “he is identical with George”). But sometimes it has an entirely different meaning. When you say “Smith is tall,” you are not saying that Smith is *identical* with tallness; you are not making an identity statement. So oftentimes “is” (and “are,” “am,” etc.) doesn’t denote the relation of identity and instead denotes the relation of *having* (as in having an

attribute). In other words, "is" sometimes has a *predicative* meaning: it is used to predicate a property of an individual. (To predicate a property of an individual is to attribute that property to that individual.) But in this context, we will focus on the concept of identity, not on the concept of attribute possession.

That said, it should be pointed that, although it is generally clear what is meant by statements involving the term "is" (or some other form of the verb "to be"), there are cases where it is functioning in *both* ways—where it *both* indicates identity *and* attributes a property to somebody (or something). If you say "Smith is the guy who washes my windows," that seems *both* to identify Smith with a certain person *and* to attribute various properties to him (e.g., the property of a window washer). We will find that these borderline cases expose important facts about the concept of identity.

1.2 Metaphorical uses of identity-related expressions

Identity statements are often used in a loose, figurative fashion. One often says things like "George isn't the person he used to be—having money has changed him" and "I'm not myself today." These statements are not meant literally. Although having money may have changed George, since he now has unsavory character traits that he used not to have, George hasn't *literally* become a person distinct from the person he used to be. It isn't as though George's brain was replaced with that of an alien. Similarly, when one says "I'm not myself today," one isn't saying that one is literally not identical with oneself—such a statement would be nonsensical. One is merely saying that, because of fatigue (or some such), one isn't doing as well as one would like (or some such). Henceforth, we will disregard non-literal sentences like "George isn't himself today" and will concern ourselves only with what is literally meant by "identical with," "identity" (etc.).

1.3 Numerical vs. qualitative identity

Terms such as "identical," "identity," etc. sometimes refer to "strict" or "numerical" identity, and sometimes refer to "qualitative" identity.

For x and y to be ***qualitatively identical*** is for them to have the very same

features (or, in any case, it's for them to be as similar to each other as is possible given the one is not the very same entity as the other). x and y can be qualitatively identical *even if they are not literally the same thing*—even if, in other words, x and y are two things, not one. If you are cloned, you and your clone are qualitatively identical, meaning that you and your clone are very similar. But there are still two of you.

In philosophy, we are concerned with the other kind of identity, to which we will henceforth refer as ***numerical identity.*** This is the kind of identity involved in statements such as "water is identical with H_2O." This doesn't say that water and H_2O are merely very similar. It says that water and H_2O are but *one* thing. Other examples of statements affirming strict identity are "the square of 4 is identical with 16," "Clark Kent is Superman," and "light is electromagnetic radiation."

1.3.1 Numerical identity entails qualitative identity, but not *vice versa*

If x and y are literally the same thing, then x resembles y to the highest degree possible—after all, x cannot, under that circumstance, have any attribute that y lacks. But x and y can be qualitatively identical—that is, they can resemble each other a great deal, without being numerically identical, as we noted when discussing the possibility of cloning.

1.3.2 Correction: Numerical identity *does* entail qualitative identity and *vice versa*

Strictly speaking, qualitative identity entails numerical identity. If x and y are truly *exactly* the same—if there is *no* characteristic that the one has that the other lacks—then they must be the same. For, at any given time, they must occupy the same place and must be doing the same things.

It can be proven formally that qualitative identity entails numerical identity. If x and y are numerically identical, then, by hypothesis, neither has any property lacked by the other. x has the property of being identical with x. Therefore, y has the property of being identical with x as well. So y = x.

The principle that genuine qualitative identity entails numerical identity

is known as the ***Identity of Indiscernibles.*** This principle is to be contrasted with the Indiscernibility of Identicals, better known as ***Leibniz's Law,*** according to which x and y, if numerically identical, are also qualitatively identical.

2.0 Objectual identity

Sometimes identity statements concern *people* or other animate beings, and sometimes they concern inanimate objects. Examples of the first kind are "Clark Kent is Superman" and "Einstein is the physicist who invented Relativity Theory." Examples of the second kind are "that's the shirt I just bought," "the car sitting in my driveway today is the very one I drove back in the '70s," and "the liquid in your glass is the water that you poured into it a moment ago—I haven't switched it with poison."

If an identity statement concerns an inanimate object (or substance), then it affirms a case of *objectual* identity. "Objectual" is the adjective form of "object." If an identity statement concerns people or other animate beings (e.g., dogs, martians), then it concerns *personal* identity. (The term "personal" is used even if the animate being in question is not a person—even if it is, for example, a dog.)

2.1 Objectual identity more fundamental than personal identity

My guess is that most people would find the topic of personal identity more interesting than that of objectual identity. A bit of autobiography may make it clear why I say this.

What is it for me—a middle-aged professor who rarely drives over the speed-limit—to be the same person as the guy who, 20 years ago, was out partying every night? I have changed so much in the meantime that it almost defies comprehension how *I* could ever have been some young roustabout. My musical horizons, for example, extend no further than the easy-listening channel on my radio, and I no longer have any tolerance for the cacophony of "rock music," which used to be my life's blood. I spend my evenings either working or reading. And when I'm not engaged in one of those two activities, it's probably because I am hobbling (or, more likely, using my "Rascal"—a

motor-powered personal locomotion device for the decrepit) at a rate of around 1/2 mph, from my desk to the easy chair in which I watch television shows targeted for senior citizens (e.g., "Murder She Wrote"). Not a single one of these statements would have applied to the person I was when I was your age.

This is not to mention that my once supple joints are frozen with arthritis, and my once pink and healthy vital tissues are now an oxygen-starved gray, thanks to the scourges to one's health brought on by middle age. In high school I was a creditable athlete, playing on three different teams. These days, even a game of shuffleboard pushes me to my limits; and I'm even considering giving that up, since there's no Rascal-parking anywhere near the local shuffleboard court.

It thus defies comprehension how, given my current condition, I could possibly be identical with the party animal of two decades ago. This problem, and others like it, are very interesting—much more interesting than the question of why the chair I'm sitting on now is identical with the chair I was sitting 10 minutes ago, when I first started writing this section. In other words, personal identity seems more interesting than objectual identity.

The problem is that personal identity must be understood *in terms of* objectual identity. Why is this? There are two dimensions to the concept of personal identity, the reason being that people have both minds and bodies, and statements about personal identity often therefore depend for their truth on both bodily and mental facts. By contrast, objects (chairs, rocks, etc.) don't have minds—they are pure body. So identity statements concerning them (in other words, objectual identity statements) depend for their truth only on physical facts. Identity statements that concern human or animal *bodies*—for example, "this is the hand that signed that piece of paper"—concern *objectual* identity. And the reason, of course, is that bodies are *objects*—like rocks and trees. They are not *subjects,* like you and me. Statements about personal identity are often concerned *both* with physical *and* with mental facts. Since they concern physical facts and objects, such statements are concerned, at least in part, with the concept of objectual identity. One must therefore understand objectual identity *before* one can understand personal identity, just as one must understand single-variable calculus before one can understand multivariable calculus.

2.2 One kind of objectual identity: diachronic objectual identity

A story will help clarify the nature of statements concerning objectual identity. At 11:00 A.M. I make myself a sandwich. But I don't eat it; I just place it on the kitchen counter. Let X be the object that at *11:00 A.M.* I place on the kitchen counter. At 1:00 P.M. I see my roommate, Smith, holding and apparently intending to consume the sandwich that I made. (There is no doubt that it's the same sandwich.) Let Y be the object that *at 1:00 P.M.* Smith is holding.

Under these circumstances,
the statement:

(*) X is identical with Y

is true. But exactly what does this mean? X doesn't have the same properties as Y. First of all, X is on the kitchen counter, whereas Y is in Smith's hands. X, having just been made out of refrigerator-stored ingredients, is nice and cool. Y, having been sitting on the counter for two hours is warm and yucky. Given how different X and Y are, what sense can it make to say that they are *identical?*

Here we have to take a brief detour through a bit of physics and metaphysics. X is not a simple entity. It has various parts—it consists of two slices of bread, a leaf of lettuce, etc. But each of those parts consists, ultimately, of a vast swarm of incessant displacements of mass-energy. Given any macroscopic physical object—a rock, a bar of soap, a plant—not a single micro-constituent of that object will stay in the same place for any amount of time. Given any one of the atoms composing any macroscopic object, that atom will move (it will rotate, vibrate, and altogether change its location) if you give it so much as a trillionth of a second. And all the objects that we deal with in everyday life—our cars, our clothes, our toothbrushes—are composed of these incessant movements of microparticles. So, really, all of those objects are collections of *events*. The *seeming* changelessness of the poster on your wall is an illusion: at any given time, that poster consists of nothing but movements and exchanges of mass and energy.

We must make two final points before we can say what (*) means. First, in some cases, events are causally related to one another, and in some cases they're not. I brushed my teeth this morning and you put on your shoes. The one event didn't cause the other. But my brushing my teeth did cause various particles on my teeth to dissolve or change location.

When events are causally related to each other, they satisfy one of two conditions: *either* they are adjacent in both space and time *or,* if they are not adjacent, they are connected by a series of events each of whose links *is* adjacent with at least one other event in that same series. The event of my dragging the toothbrush across the bit of cereal stuck to my tooth occupied the same approximate place as the event of the bit of cereal's leaving my tooth, and the first event came right before the second. So, in this case, the cause (my moving the toothbrush over my tooth) and the effect (the bit of cereal's being dislodged from my tooth) are adjacent in space and time.

Not all causes are related to effects in such a direct way. I turn on a flashlight by pushing a button on it. The event of my pushing this button causes a light beam to travel from my flashlight to some distant planet. So one event (my pushing a certain button) causes another event (a certain spot's appearing on an otherwise dark, planetary surface). These two events are separated in space and also in time (it takes time for light to travel, and the larger the distance, the more time is needed). Nonetheless, the cause and the effect are *connected* by a series of events each of which *is* adjacent, in space and time, with another event in that same series. Let t be some period before the light beam reaches the planet but after it has left my flashlight. Suppose that, at time t, the tip of the light beam is in place p. The events—the swarm of photons—in place p at time t cause there to be similar photon-swarms at the place right next to p at the time right after t. So if p* is some place right next to p, and t* is some time right after t, then the photon-swarms in p at t cause there to be similar photon-swarms in p* at t*. In their turn, *those* photon-swarms—the ones in p* at t*—cause there to be other, similar photon-swarms in some third place (one that is right next to p*) at some third time (one that is right after t*). So even though my pushing the button is remote, in both space and time, from the resulting spot of light on the distant planet, those two events are connected by a *continuous* series of events each of whose installments is caused by its predecessor and causes its successor.

Any such series is known as a "causal sequence." Light beams are causal sequences; but so, for the reasons given earlier, is *every* object. X (the cool 11:00 A.M. sandwich on the counter) is part of the same causal sequence as Y (the warm 1:00 P.M. sandwich in Smith's hands).

In light of these points, we can say what (*) means. (*) says that the events that constitute X and Y are both parts of some one causal sequence. To generalize this point, many statements of the form *a is identical with b* say that some object is causally connected in a certain way with some other object.

Here's another example. Let B be the person who is, at this very moment, writing this handout; and let A be the earlier described wild man from many years ago. The statement *A is identical with B* is true because B resulted, by a series of continuous, causally connected events that were initiated by events constituting A.

To sum up, in some cases, *x is identical with y* says that x and y are different installments in some one causal sequence.

2.3 Another kind of objectual identity: synchronic objectual identity

But sometimes *x is identical with y* has a very different meaning. Suppose that the world's most valuable diamond is the only diamond in the Hawthorn Museum. In that case, the following identity statement is true:

(#) "the world's most valuable diamond is identical with the only diamond in the Hawthorn Museum."

But # isn't saying anything about causal sequences. This becomes clear when we replace # with the (let us suppose) true statement:

(#t) "the most valuable diamond in the world *at time t* is identical with the only diamond that is in the Hawthorn Museum *at time t.*"

#t concerns a single instant. It doesn't concern a stretch of time; and it therefore doesn't concern causal sequences, since those, by definition, take

up more than an instant of time. What # and #t are saying is that there is some one object that has two different properties. So # is saying that:

(##) there exists a certain object that is a diamond that is more valuable than any other diamond in the world and that, at this time, is in a certain museum (the Hawthorn museum) that contains no other diamonds.

Thus, #t is saying that there is some property P and some property P* such that some one thing uniquely has P and also uniquely has P*. Many identity statements are of this kind. Consider:

(SC) "Superman is Clark Kent."

This doesn't tell you that two different things are one. Two things can't be one. If SC said otherwise, it would not only be false but absurd. SC makes a non-absurd statement about properties. It says of the property of being a super-hero who lives in Metropolis and has a domicile on the North pole (etc.) that exactly one thing has that property; and it says that anything that has that property also uniquely has the property of being a wimpy reporter at the Daily Bugle who is the butt of Lois Lane's derision (etc.).

Similar remarks hold of:

(WH) "water is H_2O (or, rather, any body of water is identical with some collection of H_2O molecules)."

WH doesn't make the absurd statement that two different things—water, on the one hand, and H_2O, on the other—are the same thing. WH makes a statement about properties. One of those properties is that of being the liquid that we drink, bathe in, etc. The other is the property of having a certain chemical composition. WH says that the first property is uniquely instantiated and, in addition, that anything instantiating that property also uniquely instantiates the second.

To sum up: $a = b$ says one of two things: (i) there is some causal sequence of which a and b are different installments or (ii) there is some one

thing that has a number of different properties.

3.0 Personal identity

Suppose that each of "a" and "b" is an expression referring to a person. In that case, of course, "a = b" is an identity statement—but it is one that concerns *personal* identity, and not merely objectual identity. Nonetheless, the principles in terms of which objectual identity is to be understood are pretty much the same as those in terms of which objectual identity is to be understood. For even when it is a statement about personal identity, "a = b" says either (i) there is some causal sequence of which a and b are different installments or (ii) that there are properties P and P* such that each is uniquely instantiated and such that nothing instantiating the one doesn't instantiate the other.

Consider the statement:

(IB) "the inventor of bifocals is identical with the first postmaster general."

IB says that there is some *one* person (Benjamin Franklin, as it happens) who both invented bifocals *and* who was also a postmaster general before anyone else was. (To avoid philosophically empty complications, we'll assume that BF is still with us.) So IB says that somebody has a number of different properties. So, in this respect, IB is SC, #t, and WH.

Now consider the statement:

(OF) "the old fogy who, at this very moment is writing this chapter (which no one will ever read and whose comedic excellence will therefore go unappreciated) is identical with the guy who, many moons ago, drove from Los Angeles to New York.

OF says that one set of events—the events constituting a person who, right now, is writing about personal identity—is an installment in a causal sequence of which an earlier installment was some set of events that constituted some person's driving from Los Angeles to New York. So OF is like our earlier statement (concerning the sandwich):

(*) X is identical with Y.

Each of OF and * affirms that two different, non-simultaneous events (or sets of events) are installments in some one causal sequence.

So the logical structure of statements concerning personal identity is not different from the logical structure of statements of objectual personal identity.

3.1 Personal identity: the mentalistic conception

Human beings have both physical and mental (i.e., psychological) characteristics. I have a certain weight. This is a physical characteristic. I have certain beliefs and interests. These are psychological characteristics. Human existence thus has a dual nature.

It is vigorously debated how the mental and the physical are related to each other. A related, equally vigorous debate concerns the nature of the events that must occur if a person is to continue to exist.

A story will clarify these points. Smith is in an accident and all of his brain is destroyed, except for the parts that regulate involuntary functions (e.g., breathing, digestion, etc.). So Smith is "brain-dead." Question: Does Smith still exist? My own view is "no." *Legaly* speaking, of course, Smith exists as long as his heart is beating. But legal categories are of little use when it comes to settling philosophical issues[289], and it seems to me that Smith went out of existence the moment his mind did.

Why do I have this view? Because I have what is referred to as a "mentalistic" conception of personal identity. This is the view that you are your mind. People are identical with mental entities of some kind. This doesn't mean that people aren't also identical with physical entities. It means that *if* people are identical with physical entities (e.g., with their brains or certain parts thereof), it is only because those entities are themselves identical with mental entities. If, for example, it turns out that the mental entities with which people are identical are also brain-structures or patterns of brain activity, what follows, the mentalist believes, is *not* that people are non-mental entities but that people *are* mental entities and that mental entities, in their turn, are also physical entities.

According to mentalism, the statement "I weigh 200 pounds" really means "my *body* weighs 200 pounds" and therefore doesn't really say how much *I* weigh. And given that I am identical with purely mental entities, I weigh no more than they do—which isn't very much, since they weigh nothing if they're non-physical, and next to nothing if they *are* physical (the reason being that, if they're physical, they're patterns of neural activity that have little or no mass).

3.2 An important fact about the mentalistic conception: a materialist can have a mentalistic conception of personal identity

One can believe that mental events are identical with certain kinds of *physical* events and *still* have a mentalistic conception of personal identity. So, for example, if you believe that feelings, intentions, and mental events in general are patterns of neural activity, *and* you believe that a body that isn't characterized by such patterns of activity doesn't house a person, then you have a mentalistic conception of personal identity, *even though* you believe that mind facts are identical with body facts. Most contemporary mentalists fall into this category, and I myself am inclined, with some reservations, to think that it is the right view.

Thus, even if you are what is known as a "materialist"—even if, in other words, you believe that mental events *are* physical events of some kind—you can still have a mentalistic conception of personality identity. This will be the case if you believe that a body that doesn't involve the relevant kind of physical events—the patterns of neural activity that are identical with beliefs, feelings, etc.—fails to be a person.

4.0 Personal identity: the non-mentalistic (or physicalistic) conception

Many philosophers these days would disagree with much of what I've just said. For these days, non-mentalistic, "bodily" conceptions of personal identity are dominant. So many contemporary philosophers would say that brain-dead Smith still exists.

Why say this? There are some obvious reasons. First of all, we can truthfully say, while pointing to Smith's thought-free body "that *person* is brain-dead." It seems a simple fact that we can describe *people* as "brain-dead" or even as just plain dead. Of course, dead people don't have anything going on in their minds. For expository reasons, we're setting aside the hypothesis that there is life after death). But we can still correctly, and therefore meaningfully, say things like "that *person* is dead."

Further, the (brain-)dead person of whom we've been speaking is identical with some *body—not* a body, like yours or mine, that mediates thoughts and feelings, but just a plain old body.

Here's a rundown of the relevant facts:

(1) Statements such as says "that person is (brain)dead" are often true (sadly).

(2) Such statements must be about people. Otherwise, it would be inappropriate to use the word "person." When pointing at (brain)dead Smith, one would have to say "that *non*-person is dead."

(3) The (brain)dead Smiths of this world don't have psychological characteristics. Therefore, the right conclusion is:

(4) A thing can be a person without having thoughts or any other psychological attributes.

The argument just given, though accepted by many, involves a gross overestimation of what our linguistic practices can give us in the way of philosophical illumination.[290] It's true that we refer to certain (brain)dead objects as "people." But, it seems to me, that could well just be a verbal convention. We refer to certain kinds of scooter as "rascals." But they're not *really* rascals. (A rascal is a mischievous young person.) I also think that our tendency to describe (brain-)dead bodies as "people" has emotional roots. At some level, we are holding onto the idea that, deep within that ravaged body, is a mind. In any case, the present author puts no stock in arguments such as (1)–(4).

4.1 A better argument for the physicalistic view

But there is a much better argument for the physicalistic conception of

personal identity. A story will help us state this argument.

Smith is a normal adult with a vigorous mind. One day he's hit by a bus. There is a complete cessation of all mental activity. Smith's stream of mental activity has been snipped. But Smith's body doesn't die. It goes into a vegetative coma for 20 years. During that time, there is no mental activity at all in what was once Smith's brain. The brain centers regulating automatic functions (e.g., respiration, digestion) work just fine, but the parts involved in mental activity are totally inactive. No thoughts, no dreams, no sensations. Nothing. After 20 years, the body wakes up, and the hitherto inactive brain centers spring back to life. The occupant of that body is obviously in complete command of his faculties. His personality traits are just like those of the pre-accident occupant of that same body, as are the values, the memories, the cognitive aptitudes, and so on. Let X be the person who occupied that body before the accident and let Y be the person who's been occupying it since the accident.

Is X the same person as Y? Is the post-coma person Smith? Or is it somebody who is psychologically a lot like Smith, but isn't *actually* him?

Surely it really is Smith. That's what we'd say, anyway. It's very hard to believe that Y is a brand-new person (albeit one who has the psychological characteristics ordinarily had only by adults).

This is a serious problem for the mentalistic view of personal identity. According to the mentalist, Smith *is* his mind. No mind, no Smith. Whatever his mind is exactly, it would make no sense to say "his *mind* is fine—but there's not so much as a flicker of mental activity going on in that head of his." Thus, no thoughts, no mind. So if the mentalist is right "no thoughts" means "no mind"; which means "no Smith"; which means that the minute Smith's mental stream was cut off, Smith ceased to exist.[291]

4.1.1 An unsuccessful mentalistic rejoinder

"But so what?," a mentalist might ask:

> "Smith went out of existence 20 years ago. And he came back into existence a few days ago. The same happens to everyone every day. You go to sleep. During that time, except when you're dreaming, your mind is gone, and so, therefore, are you. You wake up, and your mind is back,

and so, therefore, are you. People go away. They come back. What's the problem?"

Here's the problem. Nothing that goes out of existence comes back into existence. Ever. No exceptions. Therefore, if *you* can survive interruptions of your mental existence—if *you* can awaken from a sleep, or a coma, during which there was *no mental activity at all* in your cranium—then *you* are not a stream of mental activity. And if you are not such a stream, it's a little hard to see how *you* could be your mind. So if *you* can wake up from thought-free slumber, then *you* aren't your thoughts or, therefore, your mind; and since it was your body, and your body alone, that existed uninterruptedly during sleep-time, you must be your body.

By way of anticipation, I myself reject one of the steps in this argument. Although I do think that one is one's mind, I don't think that one's mind is one's thoughts. (I believe, for reasons to be discussed, that one's mind is a structure that gives rise to thoughts but does not itself consist of them.) But right now, instead of saying why, in my view, this argument is objectionable, I want to make it clear why it is such a powerful argument.

4.1.2 An unsuccessful mentalistic rejoinder (continued)

What most people will disagree with in this argument is the assumption that *nothing that goes out of existence comes back into existence.*

"But surely that isn't right," it will be said. "What about reincarnation? Teleportation? Can't things take breaks from existence? Can't they, so to speak, go on holiday, and then come back later?"

No. A story will make this clear. Using your laser gun, you vaporize some rock. So what was once a very sturdy, solid rock has now been dispersed into various atoms. But you manage, through some miracle of technology, to locate all of these atoms, scattered throughout space though they are, and then to reassemble them into a rock that is qualitatively just like the rock you vaporized. Question: is the new rock numerically identical with the one you vaporized? In other words, is it *the very same rock?* Or, instead, is it a distinct rock that is *extremely similar* to the old one?

Answer: it's a distinct rock that's very similar to the old one. They are, apart from their spatiotemporal properties, qualitatively but not numerically identical.

Another, similar story will reinforce this conclusion. Rock R is vaporized —but not by a human or anything else that can act deliberately. The blind, uncaring hand of nature is what does the vaporizing. R's atoms are dispersed through the cosmos. But, by some extraordinary coincidence, they all reconvene a million years later and they attach themselves to one another, thereby forming a new rock. Let R* be this new rock. Of course, R* consists of exactly the same particles as R. Moreover, R*'s structure is exactly like R's. Finally, each of the atoms composing R* has exactly the same role within R*'s structure that it had within R's structure. So

R

and R* are, apart from their spatiotemporal differences, qualitatively identical. *But surely they're not numerically identical.*

4.2 Continuity the essence of diachronic identity

Continuity is necessary for identity. It's only a slight exaggeration to say that it *is* identity. If, on the other side of the world, atoms were to collect and form a creature exactly like you—same physical and psychological characteristics —it wouldn't be you. (You're here. That other person is over there.) If you predecease this doppelganger, he obviously won't *become* you; he won't take over your identity any more than he did when you were alive. It's irrelevant that *qualitatively* he's just like you. He doesn't pick up where you left off. Supposing that you died at time t, pre-t you doesn't flow uninterruptedly into post-t him.

Let's change the story a bit. Right now, there's no doppelganger of you. You pass on (sadly). So, for a period, there's no one in existence who has exactly your characteristics. But, by some extraordinary coincidence, various atoms come together and form a creature qualitatively just like you, who we'll call "new-you." Is new-you you? No. Does it make any difference whether the atoms composing you are the ones composing new-you? No. Once you're gone, you're gone.

4.3 Why the concept of teleportation is an incoherent one

On *Star Trek,* people were routinely teleported. Any case of teleportation has three phases. Phase 1: the disintegration phase. The teleportation candidate is decomposed into small particles. Phase 2: the movement phase. Those particles are moved at or near the speed of light to another place. Phase 3: the reintegration phase. At the new place, those particles are reassembled into a being that is qualitatively just like the one that, moments earlier, was disintegrated. So the Phase 3 entity is exactly similar to the Phase 1 entity.

Is the Phase 3 entity numerically identical with the Phase 1 entity? No. Once Captain Kirk has been disintegrated, he's gone. What shows up on the Enterprise isn't him; it is, at most, a doppelganger.

4.3.1 Structure-preserving continuity, as opposed to mere continuity, as the essence of identity

My students often object to this. Here's their argument: "Teleporting Ct. Kirk (or whoever) involves transforming his energy into a new form, and then, at the target location, converting that same energy back into the old form." I agree with the first part: Teleporting Ct. Kirk would indeed involve transforming his energy into a new form. But so would dropping a one million ton weight on him or dispersing his atoms across the cosmos. "Same energy" doesn't mean "same object." I've been around for three-and-a-half decades. My energy has been around a lot longer than that. Thirty five years ago that energy came to be embodied in a structure that, through a series of piecemeal, continuous, and *structure-preserving* changes eventuated in the structure that is writing this sentence.

The changes in question had to be "structure-preserving" because, as we'll see in Sections 7.0–7.5, a continuous causal process that mediates drastic structural changes may not be enough for identity. The person currently typing this sentence arose, by a continuous series of changes, from a certain tadpole-like creature. But, structurally, I'm sufficiently different from that tadpole that it would be a stretch to say that I used to *be* it. But I'm structurally enough like the blobby infant in the picture that, given my causal

connectedness to him, I can say that I used to *be* him.

4.4 Once gone, always gone

Let's map these points onto our points about teleportation. The Phase 2 atom-scatter has *no* meaningful structural resemblance to the Phase 3 human who, when these atoms are integrated, appears on the deck of the Enterprise. The degree of structural resemblance between me and the zygote that eventually turned into me is far greater than the structural resemblance between the Phase 2 atom-scatter and the Phase 3 Enterprise occupant. We thus have no choice but to say that Kirk was obliterated the moment his atoms were scattered. Whatever it is that "beams aboard" isn't Kirk: it is, at best, a good imitation of him.

4.5 The physicalist's high-card: the ice-man

Of course, certain *aspects* of existence can be interrupted. You can temporarily shelve your career as a hockey player. But *you* cannot be shelved. "But what if I'm frozen and kept in ice for a thousand years, and then thawed out and reanimated? Would that be me?" Yes, it would (or so I think). But what that means is that, when you were on ice, *you* were still in existence. If, instead of being frozen, you had been disintegrated and, a thousand years later, the atoms composing you were reassembled into a creature just like you, we would *not* say that *you* had come back; we'd say that, although you were gone, somebody just like you popped into existence.

So, returning to our ice-man story, if post-thaw you is identical with pre-thaw you, that's because the years of frozen stillness didn't *interrupt* your existence. "But," you protest, "my mind was frozen during that time. No dreams. No thoughts. Nothing (in the way of mental activity). So, clearly, my existence *was* interrupted."

Your *mental* life was interrupted—but not your existence *tout court.* "But," you respond, "you agreed that the post-thaw person is really *me*—that the pre-thaw person and the post-thaw person are identical, just as the person typing this sentence is identical with the person who typed the sentence before it."

Yes, I do agree. "But that makes no sense, given the reasonable

supposition that (i) I *am* my mind and given that (ii) my mental existence was severed." Yes—it makes no sense *given* those two assumptions. (By way of anticipation, we'll find that (ii) is false.)

But it *does* make sense given a different supposition, namely, that you are your *body*. Your body remained in existence during the freeze years. There was no interruption. To be sure, certain forms of *activity* were interrupted. Everything associated with mentation was on hold. But it's a datum that your body was around—that, in fact, it was kept more intact during that period than most living organisms are kept for a period of more than a day. But, evidently, *you* were around.

Bottom line: a sentient creature can survive interruptions of its mental existence, but not its physical existence. Interrupted existence is non-existence. So a sentient creature is identical with its body.

That is the main argument on behalf of the physicalistic theory of personal identity. And it's a good one. To refute it, the mentalist would have to show *either* (a) that existence can be interrupted *or* (b) that cessation of mental *activity* isn't the same thing as cessation of mind. (a) is a non-starter; establishing it would be like trying to establish that squares could have five sides. But (b) is quite defensible.

4.6 A defense of (b): the mentalist strikes back

People have relatively stable personality traits and beliefs. Yesterday I believed that snow is white. I still believe it, and I'll believe it tomorrow. This holds of countless other beliefs of mine, as well as attitudes, aspirations, values, fears, etc.

People change, of course. The deepest changes happen in one's early years. But the change never entirely stops, even though it does become increasingly internal to established structures.

Throughout all this change, however, there is a great deal of stability in one's beliefs, attitudes, etc., and also in one's basic personality architecture. And the change itself is, in is broader outlines, continuous. It doesn't typically happen in fits and starts; it is typically more evolutionary than revolutionary. Of course, there are revolutions—sudden epiphanies or transformations. But they are internal to a larger, continuous structure whose

integrity they don't undermine.

Where there is regularity, there is typically an underlying structure; and where there is irregularity, there is typically a lack of structure. A person who has a job tends to act in the same ways over and over. This is because his activities must be internal to a certain organizational framework. By the same token, a person who doesn't have a job doesn't have to behave in those ways. Other things being equal, a person without a job behaves more erratically than one who doesn't.

In light of this, let's consider some humdrum facts about my psychology. When I woke up today, I believed that 1 + 1 = 2. When I went to sleep last night, I believed it. I obviously wasn't thinking *1 + 1 = 2* the entire time I was asleep. (And even if, by some remote chance, I was, there are many other beliefs that I have that weren't coursing through my psyche during sleep. So just let my belief that 1 + 1 = 2 represent any one of those.) If that belief simply ceased to exist when, upon my falling asleep, it became inactive, it would be a miracle that, on waking up, I resumed having that belief. That would be a causal anomaly, a case of creation *ex nihilo.* That belief, therefore, must have its basis in some kind of *structure* that endures even when it isn't active.

In fact, beliefs *are* structures. Beliefs must be distinguished from the use to which they're put. At any given time, most of your beliefs are inactive. You're not drawing on them, arguing for them, reconsidering them, or emoting about them. But during these lulls the beliefs themselves no more cease to exist than a sheriff's gun ceases to exist when he's not firing it.

Beliefs are not happenings. They are not, as philosophers say, "occurrent." A searing pain is occurrent. So is any sense-perception or act of thought, whether conscious or unconscious. But beliefs aren't processes. Nor are attitudes or intentions or aspirations.

No individual belief changes. For a person's beliefs to change is for that person to shed some old beliefs or acquire new ones. I used to be convinced that Platonism is false; now I'm convinced that it's true. But that change was a case of one belief's being obliterated and replaced by a different one; it wasn't a change that was internal to the persistence of some one belief. Whenever a belief is "modified," what's going on is that it's being replaced

with another, possibly related belief. And it isn't the *belief* itself that is occurrent: beliefs are always stable. What is occurrent is the process whereby the one belief is replaced with the other.

The same is true of attitudes, personality traits, and intentions. What is occurrent isn't the intention itself; it's the process of acquiring it or of dumping it (and replacing it with another).

One's beliefs, attitudes, etc., are an important part of one's mind. Since these things are structures, it immediately follows that there is *much* to one's mind that isn't a process of any kind. What about the occurrent processes are associated with mind? The pains, the pleasures, the sense-perceptions, the waves of doubt, the conscious acts of deliberation? These, it seems to me, are mental *actions* and are thus no more constitutive of your mind than a bodily action (e.g., one's jumping over a fence) is constitutive of your body.

One's mind isn't one's stream of mentation; it is the structure that mediates this mentation. We saw reason to believe this earlier. So the slogan "no thought, no mind" is false. The right slogan is: "no thought-enabling structure, no mind."

This slogan suggests another one: "mind = thought-enabling structure (as opposed to stream of thought)." When one is in a deep sleep or even cryogenically frozen, the structures responsible for thought are inactive, but they don't cease to exist on that account. (An elevator that isn't being used still exists.)

4.7 An objection

Let's consider an object to the argument just presented:

> "But aren't those structures brain-structures? Ice-man survives his thousand-year ice coma only because those brain-structures remain intact. Those structures are obviously physical; they're part of one's body, like one's heart or gall-bladder. So your theory of personal identity (viz. one's self is one's mind and one's mind consists of the structures underlying one's thoughts) collapses into a physicalist, and therefore a non-mentalistic, analysis of personal identity."

Mentalists usually hold that mental entities are physical entities. A mentalist who holds that thought-structures are brain-structures will hold that, when those brain-structures go, so will the corresponding person, the reason being that those brain-structures are identical with that person's mind. The physicalist holds that, no matter what happens to the parts of brain mediating sentience, a given individual survives *provided* that his heart, lungs, gall-bladder, etc. are still working. His reason is that one's self is one's body and it's a fact that one's body can remain alive, and certainly exist in some form or other, well after one's mind is gone. The difference between physicalist and mentalists isn't that the former do, whereas the latter do not, believe the minds are brains (or brain-structures). The real difference is this. Mentalists believe the *only* parts of the mind to be directly responsible for personal identity are the parts that are identical with one's mind. Physicalists don't think that these brain-structures are any more important to the persistence of the self than parts of the brain that have nothing to do with thought (that, for example, regulate the manufacture of hormones).

5.0 Degrees of identity

It's generally held that, for any object x and any object y, x and y are either identical or they're not; they can't be identical to a certain degree.

Is this correct? Remember that we distinguished between diachronic and synchronic identity-claims. "Barack Obama is the current U.S. President" is a synchronic identity claim, and "I [JMK in 2009] am identical with that blobby infant [in the photo from 1973]" is a diachronic identity claim. Where synchronic identity claims are concerned, there are indeed no degrees of identity. x and y are synchronically identical or they aren't—there's nothing intermediate between total synchronic identity and total synchronic distinctness.

But diachronic identity is a different matter altogether. Diachronic identity not only can come in degrees, but *must* do so. Let "x = y" be an arbitrary diachronic identity claim. So far as it's true, it's only if the following conditions are met. First, y is a causal successor of x. In other words, there is some sequence of events, such that each event in that sequence is caused to occur by its predecessor (if it has one) and causes its

successor to occur (if it has one), and such that x and y are both installments in that sequence. But for "x = y" to be true, it isn't enough that these conditions be met. If x is the particular gamete that, in due course, became me, and y is me, "x = y" isn't true.

But why not? Because x and y are *too* different. For "x = y," when taken as a diachronic identity-statement, to be warranted, it is necessary not only that y be x's successor in some causal sequence, but also that the relationship borne by x's condition with respect to y's condition be a predictively and explanatorily robust one. I am structurally *so* different from the gamete that eventuated in me, and the bridges that had to be crossed in order for that gamete to eventuate in anything at all like myself were so rickety, that the relationship borne by that gamete to my present self isn't that of young JMK to old JMK, but is more like that of ancestor to descendant. To be sure, a single entity can change a great deal. The butterfly is structurally very different from the caterpillar that became it. But the predictive and explanatory relations borne by the caterpillar with respect to the butterfly are sufficiently tight enough that it's more fruitful to see the change in question as internal to the existence of some one thing than it is to see it as a case of one thing's being replaced by some distinct thing.

Given two bodies of data, D_1 and D_2, and some event E, D_1 can be more relevant to E than D_2; D_1 can do a better job than D_2 of explaining E (if it has already happened) or of providing reasons to predict E's occurrence (if it hasn't yet happened, but will). Explanatoriness and predictiveness come in degrees. Since "x = y" is true *to the degree* that x bears the relevant explanatory and predictive relations to y, thinghood is a matter of degree. It's also a matter of context. In some explanatory contexts, it *would* be correct to say of some gamete that it is identical, in the diachronic sense, with some philosophy professor.

These points are developed in Sections 7.0–7.5 of the present chapter, which is concerned with the conditions that two different states of affairs must satisfy if they are to both be constitutive of some one entity.

5.1 Degrees of personal identity (continued)

We tend to think that *personal* identity either holds or it doesn't. "I remember *being* little Timmy. (I remember playing kickball with my friends; I remember learning how to read; etc.) If I remember being Timmy, then I am Timmy. It makes no sense to say that I might be 83% identical with so and so." On the contrary, it makes plenty of sense. True—you have various memories (of playing kickball, watching cartoons, waiting for the ice-cream truck to come to your block, etc.) But all that shows is that certain currently existing mental structures and events are derivatives of the mental structures and events associated with some body (literally, some *body*) that is related to the body you now occupy in the way that a caterpillar's body is related to that of the butterfly into which it metamorphoses.

We're dealing with relations of succession and causation—nothing more. There isn't some mysterious inner flame, some underlying substance, that remains untouched throughout it all and that somehow underlies the fact that you and the little kickball player are identical. Even if there is such an inner flame, it is inert *except* in so far as it facilitates the relevant psychological and physiological events. But so far as it does facilitate them, those events do all the work—they are what "keep you going", so to speak. The inner-flame is a free wheel and thus drops out of the picture.

Thus, so far you are identical with the little kickball player of yesteryear, it is because you and he are connected to each other by a sequence of events whose earlier members provide information about, and provide the causal basis for, its later members. But, as we said, one thing provides different amounts of information about another, and one thing can be causally responsible to different degrees for another. So identity in the diachronic sense is a matter of degree.

"If you're right," it will be asked, "why are we so convinced that diachronic identity is absolute—that I either *am,* or am not, the little kickball player?" I'm not sure that we are so convinced. What does a person even mean when he says that he *is* the little tyke in the photo? Does he *mean* that there is some underlying, unphased soul-substance that inhabits that little tyke and also inhabits him? Surely not. So far as he means anything, it can only be something that corresponds to what we've been saying—that there is a *connection,* both causal and thematic in nature, between the little tyke in the photo and the person who is now taking about that little tyke.

5.2 Time-indexed properties and Leibniz's Law

There's no doubt that ***Leibniz's Law*** (the principle that, if x is numerically identical with y, x has a given property if and only if y has that property) correctly models the logic of *synchronic* uses of expressions such as "identical with," "is" (when used non-predicatively), and so on. (Leibniz's Law is the principle that, if x is numerically identical with y, x has a given property if and only if y has that property.) Given that Barack Obama is the Commander-in-Chief, the Commander-in-Chief has a given property iff Barack Obama has that property.

But it's less clear whether Lebniz's law governs the *diachronic* use of such expressions. The orthodox position is to say that it does. But there is reason to question this orthodoxy.

Right now I'm a professor. The blobby infant in the picture is not a professor. So I have a property that it does not have. Therefore, by Leibniz's Law, I am not it. Identity-theorists say that the blobby infant and I *do* have the same properties. To make this idea work, they bring in the highly artificial notion of a "time-indexed" property. I have the property of being-a-philosophy-professor-in-2009, and so does the blobby infant. The blobby infant has the property of being-preverbal-in-1973 and so do I.[292] Therefore, the fact that, unlike my six month old self, I am not *now* a blobby infant, being instead a non-blobby professor, does not make me different from my six month old self. Given any other property that I now have that I used to lack, or *vice versa,* a similar argument shows that I have *always* had that property. For this reason, it is said, Leibniz's Law *does* govern *diachronic* uses of expressions like "identity," "is" (in the non-predicative sense), and so on.

In addition to being obviously contrived, this view has the absurd consequence that nothing changes. Whatever properties I have now, I've always had then and always will have them. In general, nothing acquires properties. It follows that nothing ever loses a property. When a person acquires the property of being able to tie his shoes, he loses the property of being incapable of tying his shoes. In general, any case of property-acquisition is a case of property-loss. So if nothing ever acquires properties,

nothing changes. Given that there is change, the no-change must be false.

In any case, we obviously don't yet *know* that the no-change view is the right one. So were we to build a theory of identity around that view, we'd be prejudging the outcome of the debate between those who accept that view and those who reject it. Good analyses tend not to prejudge open questions; and when good analyses do prejudge such questions, they tend to prejudge them in a way that is favorable to deeply rooted intuitions. The view that nothing changes is not exactly a deeply rooted intuition. So for the time being we must regard all this talk of time-indexing as nothing more than hollow trickery.

To move forward, we must say a word about the debate that rages between those who accept a doctrine known as "four-dimensionalism" (henceforth, "4-D") and those who reject that doctrine. According to 4-D, any given object is identical with a "space-time worm." A space-time worm is an entity that is spread out in both space and time. Thus, according to 4-D, you are identical with a space-time worm that comprises all of the events and states of affairs constituting you from the moment of your birth (or conception) till the moment of your demise.

Bearing this in mind, let t be any instant in time during your lifetime. If 4-D is right, some *segment* of you exists at t; but *you* do not exist at t, since you, according to 4-D, are spread out over a period of (say) 80 years. So if 4-D is right, I cannot see you or shake hands with you; I can see and shake hands only some segment of you. The thing I shake hands with, says the advocate of 4-D, is some *time-slice* of you.[293]

In my view, I *can* shake hands with people, and also see them, play tennis with them, and otherwise interact with them. I thus reject 4-D. I am not alone in this. (But I am more isolated than one would think, given how deeply rooted the presumption is that one can see people, shake hands with them, and so on.)

Another reason I reject 4-D is that I believe that people and things change. 4-D denies this. If 4-D is right, the thing that is reading this sentence right now isn't *you;* it's just some time-slice of you. And the slightly different thing that was reading the first sentence of this chapter isn't you either; it's just another time-slice of you. The same is true of the thing that decided to

buy this book a few weeks ago. According to 4-D, although the thing reading this sentence is different in some ways from the thing that was reading the first sentence of this chapter, those differences aren't expressions of the fact that some one thing has *changed.* According to 4-D, *you* don't change. For a given object to change is for the way that object is at one time to be different from the way it is at some other time. According to 4-D, for any time t, what exist at t isn't an one object, and is only some time-slice of an same object. Therefore, there are no two moments of time t and t* such that how some one object is at t is different from how that same object is at t*. Therefore, if 4-D is right, *things don't change.*

But this is absurd. Things change. I have changed, and so have you. And each of us will continue to change. 4-D is either dead wrong or it amounts to nothing more than a proposal that we use the words "thing" and "change" in non-standard ways. And even if by some remote chance it turns that things *don't* change, it would be bad methodology to make our theory of personal identity depend on that turning out to be the case. In doing so, we'd be assuming facts (or non-facts, rather) that aren't yet in evidence.

It may be that 4-D is a cogent analysis of *some* concept. But that concept isn't the one that is ordinarily associated with expressions such as "identity," "identical with," etc. The concept ordinarily associated with such statements is an incoherent one *unless* people change. The 4-dimensionalist is analyzing some artificial analogue or reconstruction of that concept; he's analyzing an artifact—a by-product of his technical trickery—that is devoid of philosophical value and is equally devoid of psychological, clinical, and practical value. And we'll see in the upcoming section that 4-D is correct only if (i) there is retroactive causation and (ii) instances of retroactive causation have absolutely no effects on anything. Each of (i) is and (ii) is a hideous assault on logic. Should it turn out that 4-D does have these consequences, there would be no good reason to accept it.

In any case, assuming that things change, the only way to go is to say that identity, in the diachronic sense, is to be understood in terms of relations of causal connectedness—that statements like "Bill is the guy in the photo who is wearing a rugby shirt" are to be analyzed in the way described earlier. Since the relevant sort of connectedness obviously comes in degrees, so does identity.

It should be pointed out that David Hume (1711–1776), whose views we'll discuss in Section 6.0, takes a similar view, and so does Derek Parfit (1942-).[294]

Finally, it should be pointed out that those who reject this view are reading metaphysics off of sentential surface structure. The expressions we use to denote the relation of diachronic identity (e.g. "is" and "is identical with") are homonyms of the expressions we use to denote the relation of synchronic identity. Given *only* that fact, along with some obscure and ambiguous intuitions about selfhood, 4-dimensionalists infer synchronic identity *is* diachronic identity—that, when I say "I am identical with the infant in the picture," the relation I am alleging to hold between myself and that blobby infant is identical with the relation that I am alleging to hold between four and the square of two when I say "2^2=4." That inference is completely spurious; and it's one that no self-respecting analytic philosopher would make, given that the essence of analytic philosophy is a stubborn refusal to take surface structure to be a foolproof guide to truth. We'll now discuss why 4-D entails (i) and (ii).

5.3 An identity-related paradox and the need for a diachronic conception of identity

We'll now discuss why 4-D entails (i) and (ii). At time t, person X_1 splits, like an amoeba, into two people, X_2 and X_3. X_2 is just like X_1—in any case, X_2 resembles X_1 as much as you, in your current condition, resemble the person you were a minute ago. And X_3 resembles X_1 to the same degree. Further, the causal connection that X_1 bears to X_2 is qualitatively identical with the causal connection that X_1 bears to X_3.

It's obvious that $X_2 \neq X_3$. After all, they're in different places. And it would obviously be arbitrary in the extreme to say that X_2, but not X_3, was X_1. Whatever claim X_2 can make to being X_1, X_3 can make precisely the same claim. David Lewis' solution is to say that X_2 and X_3 have been there all along: as soon as X_1 came into existence, so did X_2 and X_3. There were

three people rolled into one body and one mind.

Lewis' position is absurd in the extreme. Before he split into two, X_1 was no more a compendium of different people than you are. There is no relevant psychological or physiological difference between X_1, before the split, and anyone else. If X_1 had died before the split, not even a shadow of a case could be made that "he" was really a "they"—that he was really three people. So, supposing that X_1 *does* live and, in due course, splits into X_2 and X_3, his undergoing that split *retroactively* puts two extra people in X_1's body. So Lewis' solution goes through only if there is retroactive causation. That is bad enough, but it gets worse. Even though, if Lewis' solution is to go through, there must be retroactive causation, that causation must fail to make any psychological

or physiological difference to X_1 or, for that matter, to the state of the universe before X_1's split. But to be a cause *is* to create change. So Lewis's position collapses into the absurdity that there are causes that aren't causes.

5.3.1 An identity-related paradox and the need for a diachronic conception of identity (continued)

It's obvious that Lewis is jumping through hoopes. He's working with a broken concept and is trying to cover up the cracks in it by mutilating wholesome and intact concepts. But Lewis' position is *de rigueur* if one holds that the sense in which adult-JMK is "identical" with baby-JMK is identical with the sense in which 2 is

identical with 1 + 1. We'v e already seen why. X_2 is no less "identical with" X_1 than is X_3. And X_2 is no less identical with X_1 than you are identical with the person you were a minute ago. Therefore, *if* we say that you are "identical" with that person in the same sense in which 2 is identical with 1 + 1, we must say that *both* X_2 and X_3 are identical with X_1 in the very same sense. Since it's a datum that X_2 and X_3, being in different places, are distinct from each other, it follows that, just as Lewis says, the mind and body had by X_1, prior to the split, were shared by X_2 and X_3. It follows, in other words,

that X_1, X_2, and X_3 were there all along, all rolled into one package.

Given how preposterous that is, it's clear what we must say. X_2 is "identical" with X_1 in the sense that it is X_1' s successor; the same is true of X_3. And the same thing *mutatis mutandis* is therefore true of you. You are identical with your teen self in the sense that you are that entity's successor.

True—the expressions we use ("identity," "identical with," etc.) to describe the relationship between Barack Obama and the current President, or between 2 and 1 + 1 are the same ones we use to describe your relationship to your teen self. But so what? Language is mirror of thought; thought isn't always perfect. Thought is especially prone to error when it comes to emotion-laden issues, such as personal identity. Human beings don't want to believe that they are event-sequences that can become frayed, like an old rope, and dwindle into nothingness. Human beings don't want to believe that their identities are internal to emotion-driven narratives. But what we want to believe is irrelevant. What *is* relevant is that, if we take the conventional view, then we must follow Lewis in embracing the absurdity that you could be 80 different people, even though you are physiologically and psychologically just like somebody who is but one person.

6.0 Personal identity revisited: Hume's analysis

Few have done as much to clarify the structure of the psyche and nature of personal identity as David Hume (1711–1776). Hume had a mentalistic conception of personal identity. In his view, there are no people where there are no minds. But, unlike us, he held that there are no minds where there are no occurrent thoughts—where, in other words, there is no mental *activity.* In other words, he rejected contention (b).

Why does Hume hold this? Because Hume is a mentalist *who doesn't believe in the existence of structures of any kind.* Hume acknowledges that there are events; but he denies that there is anything underneath those events that guide them. There are no forces, no structures, no powers. (See Chapter 17.) And this is why, being a mentalist, he must identify the self with a stream of mental events.

There is an obvious objection to this view: "If Hume is right, then the

person who wakes up in the morning isn't identical with the person who went to sleep at night." Hume is aware of this objection, and his response to it is very profound. He says that diachronic identity is pseudo-identity. The concept expressed by the occurrence of "is identical with" in "JMK is identical with the blobby infant in the photo" simply isn't the same concept as the one expressed by its counterpart in "Barack Obama is identical with the current U.S. President." Taking it for granted that synchronic identity is a genuine form of identity, he infers, very reasonably, that diachronic identity is pseudo-identity.

6.1 A preliminary point: Hume's empiricism

Empiricism is the doctrine that everything that is known is known through sense-perception.[295] So far as empiricists grant that anything can be known that isn't directly observed and instead inferred from what is directly observed, the inference-rules involved must themselves be known from direct observation alone.

No sooner is empiricism defined than its coherence is called into question. An inference-rule is a normative truth. It tells you what you *ought* to infer from a given premise. And, as empiricists themselves point out, observation tells you what is, but it doesn't tell you what ought to be.

In any case, Hume was a strict empiricist. Unlike other self-described empiricists, he didn't help himself to rules of inference whose legitimacy doesn't have a strictly perceptual basis. This was because, unlike his empiricist brethren, he had the acuity to see that many inference-rules that seem to be empiricism-friendly ones are not.

Hume's empiricism was doubly strict. First, he had extremely conservative views as to which rules of inference had the requisite perceptual basis. Second, he had extremely conservative views about what sense-perception *directly* tells us.[296] What Hume says about personal identity must be understood in terms of the fact that he is a very strict empiricist.

6.1.2 Hume's first argument

Hume gives two arguments for his contention that selves are composed of the

mental contents that occupy them. Here is the first:

> "For my part, when I enter most intimately into what I call *myself,* I always stumble on some particular perception or other, of heat or cold, light or shade, love or hatred, pain or pleasure. I can never catch *myself* at any time without a perception, and never can observe anything but the perception."[297]

There are two ways to interpret this. On the one hand, Hume could be saying that selves don't exist. On the other, he could be saying that selves are *constituted* by the thoughts, feelings, etc., that we describe them as "having."

The right interpretation is the second one. Notice how frequently the word "I" occurs in his argument. He obviously thinks that it has a referent, and it's a truism that this referent is a person and, therefore, a self. Also, in other parts of the Treatise, he talks at length about the relation that two different mental contents must bear to each other to belong to the same mind or self. Thus, Hume's position is that, although there are selves, selves are no more distinct from the mental events that occupy them than a swarm of wasps exists independently of the wasps that compose it. For an event to *occupy* a mind is for it to *constitute* it.

The just-quoted argument of Hume's thesis, it will be noticed, is *empirical.* Hume is arguing on observational, as opposed to strictly logical or conceptual, grounds that there isn't anything that underlies his various thoughts. When he looks inward, he finds thoughts and nothing else; in particular, he doesn't find a non-thought that "supports" these various thoughts. This fact is important, as we'll now see.

6.1.3 Problems with Hume's argument

Given only that we can't observe something, it doesn't follow that it doesn't exist. There are many unobservable entities (e.g., electrons). If Freud (1915) is right, there are unobservable (unintrospectible) mental entities (e.g., unconscious urges). Saying that selves don't exist since we can't introspect them is like saying that electrons don't exist since we can't see them.

A related possibility that Hume is ignoring is that minds are *structures* of

the kind described earlier. (This oversight is probably deliberate, as we'll see in a moment.) He's also ignoring the correlative possibility that specific mental occurrences (e.g., images, tickles) are *products* of minds, and are not themselves *constitutive* of them.

In ignoring these possibilities, Hume is stacking the deck in his favor. We are pretheoretically inclined to see minds as being to specific tickles, itches etc., what car factories are to cars—as being the things that give rise to them and, therefore, aren't constituted by them. In any case, we're pretheoretically much more inclined to see minds in this way than we are to see them in Hume's way, viz. as sequences of mental occurrences, none of which has any moorings in any structure and each of which, therefore, just happens. And in simply dismissing that possibility out of hand, Hume is question-beggingly assuming the falsehood of what is probably the most plausible rival to his own view. In any case, supposing, as Hume does, that such structures aren't known through introspection, it isn't necessary to conclude that they don't exist; for it would be equally reasonable, if not considerably more reasonable, to conclude that introspection is a limited medium.

6.1.4 Why does Hume deny that there are mental structures?

Hume is a hardcore empiricist. You know it only if you (i) observe it or (ii) observe direct evidence for it. (x is "direct" evidence of y if, given no inference rules whose legitimacy cannot be established strictly on the basis of observation, y's existence can be inferred from x's.) Hume believes that mental structures cannot be known in either of these two ways. (He's right.) So they don't exist as far as he's concerned.

Hume doesn't believe in structures of *any* kind. A structure is something that underlies events and guides them. Hume doesn't believe that anything guides anything. He doesn't believe that anything *makes* anything happen or otherwise *influences* anything. In his view, words like "force," "power," "coercion," "guide," and "regulate" are meaningless. He holds this because, so he believes, forces cannot be observed and neither can direct evidence of them. (Hume's views on causality are the subject of Chapter 17.)

Thus, Hume's empiricism blocks him from granting the existence of

mental structures, given that such structures can't be known (para)perceptually. Second, given Hume's very reasonable view that one can't observe forces or any direct evidence of them, his empiricism prohibits him from granting the existence of structures of *any* kind. Thus, Hume's empiricism makes it doubly impossible for him to grant that there is anything to one's mind other than fleeting, conscious events. (In Chapter 18, we'll discuss that, contrary to what Hume holds, forces can sometimes be directly observed.)

6.1.5 Problems with Hume's argument (continued)

In an attempt to show that there isn't anything to one's mind other than the various specific mental occurrences that it houses. Hume says that, when he looks inward, he doesn't find an inner flame; all he finds, he says, are specific images, tickles, etc.

But unless this inner flame is *itself* a fleeting mental occurrence—unless it's the sort of thing that can be disclosed to one in the form of an introspectible image or sensation—nothing significant follows from the fact that, when he introspects, he doesn't find this inner flame. So for Hume's argument to go through, it must be assumed that the self, if existent, is to be known to one in the same way as tickles and itches. It must be assumed, in other words, that the self is a fleeting conscious occurrence. But that's the very thing Hume is trying to show; and, were Hume to assume it, his argument would be viciously circular.

6.1.6 The problems with Hume's argument is reflections of his hardline empiricism

It is through images that our perceptual modalities (sight, touch, etc.) tell us about the world. (Not all images are visual; there are acoustical images, tactile images, etc.) Your seeing a tree consists in your having an image of that tree; your feeling the tree with your hand consists in your having a tactile image of that tree; and so on. There's no perception where there's no image. It follows that, if introspection is an inwardly directed perceptual modality, it tells one nothing when it isn't giving one an image.

Hume says that, when he looks inward, he encounters various

"perceptions" (images), but no self. On this basis, he says that there is no self. But the right conclusion is that, *if* there is a self, it isn't disclosed to one in the form of an image.

Supposing that it exists, there are, ultimately, two possibilities as to what the self is: it could be (i) one's center of awareness or (ii) some structure underlying the events in one's mind and guiding them. (It may be both, given that (i) and (ii) aren't mutually exclusive.) If this is right, the self cannot be known in a way that can be modeled on sense-perception, and there is thus no merit to Hume's paraperceptual demonstration of the non-existence of the self.

It is uncontroversial that we are directly, non-inferentially aware of many mental entities that we *don't* know of in some image-mediated, perception-like way. I know I believe that 1 + 1 = 2. It's obviously not

through *inference* that I know this—I know it directly. And it obviously isn't through sense-perception that I know it. Finally, it isn't by virtue of my experiencing, or being aware of, some *mental image* that I know it. *But surely I know it*. So I know it directly. It's a pretheoretic datum—not the result of an intellectual working over of raw data. No—it *is* one of the raw data.

The same is true of many other mental contents of mine. I know that I intend to write about causation later today. It obviously isn't through *inference* that I know this—it isn't because I psychoanalyzed myself or engaged in any other such artifices. And it obviously isn't through sense perception (sight, hearing, etc.) that I know it. Nor is it through the inwardly directed para-perception that Hume describes. But I know it. I know it in the same way that I know that 1 + 1 = 2, that I regret not calling my mother on her birthday, that I plan on improving my skills as pianist, etc.

I know these things through some faculty that gives me direct, non-inferential knowledge of at least some facts about my own mind. Were Hume to deny the existence of this faculty, he'd be butchering obvious facts to make his argument go through and his position would be a most unempirical one. It is thus an empirical fact that not every fact about one's mind that one is non-inferentially aware of is disclosed to one in the form of an image; and it's thus an empirical fact that not all knowledge is arrived at through perception or through anything at all like it. It is thus an empirical fact that

empiricism is wrong!

6.2 Hume's second argument: a preliminary point

Hume's first argument is presented as an empirical argument. He says that his observations yield no evidence that there is some central fire around which his various thoughts orbit. His wording, and the wording that it is natural to use when paraphrasing his argument, suggests that Hume is trying to establish the non-existence of the self on grounds comparable to those on which an astronomer would try to establish the non-existence of a certain planet. "If I had evidence, I'd believe. But I don't, so I don't."

But so far as self-observation yields *any* relevant evidence, that evidence seems to suggest that *there does exist* within oneself a center of awareness. When I self-observe, I *do* feel that there is a core-being around which everything else orbits.

This feeling of mine is by no means probative. I don't know that it's accurate. I don't know that it's a *correct* observation. In fact, I don't know that it's an *observation* at all. It may be that, in having this feeling, I'm experiencing as an empirical datum what is in fact a projection into the observational arena of my acceptance of an erroneous, theory-driven *interpretation* of the relevant data. This is Hume's position. He explicitly says that those who claim to be aware of an inner flame are either lying or they're in the grips of some preexisting, empirically unfounded theory of what the self *must* be and that, if they're observing anything, it's only a projection of their acceptance of that bad theory.

The problem is that, even if, as Hume claims, my feeling that there's an inner flame is an illusion powered by a very wrong theory, it's a datum that this *feeling* exists. And it *isn't* a datum, though it may be a truth, that this feeling is an illusion. So, even though Hume's opponent may be wrong, it's *Hume,* not his opponent, who's making the theoretical point. This suggests that he has some theory in mind, his acceptance of which, and not some body of observations, led him to accept the "no inner flame" view.

6.2.1 Hume's second argument (verbatim)

And, in fact, Hume does have a theory:

> When my perceptions are remov'd for any time, as by sound sleep; so long am I insensible of *myself,* and may truly be said not to exist. And were all my perceptions remov'd by death, and cou'd I neither think, nor feel, nor see, nor love, nor hate after the dissolution of my body, I should be entirely annihilated, nor do I conceive what is farther requisite to make me a perfect non-entity.[298]

This argument, unlike the first, isn't empirical at all. This time, Hume is trying to show on purely conceptual grounds that there is nothing underlying the various events occupying a person's mind. He's arguing that the very *idea* of such a thing is an incoherent one, like that of a square circle. Why is this idea incoherent? Because, Hume says, it has the, in his view, patently absurd consequence that there could be thought-free minds.

6.2.2 Evaluating this argument: Hume's failure to distinguish between two very different meanings of the word "thought"

Hume is right that *if* there is some inner flame *underlying* one's thoughts, and therefore distinct from them, it *does* follow that one's self could exist in a thought-free condition. But Hume is wrong to say that a thought-free self is an absurdity. The reason Hume *thinks* it absurd is that he fails to distinguish two very different meanings of the word "thought."

Sometimes the word "thought" refers to mental occurrences—to what are now known as *occurrent* mental phenomena. An occurrent mental phenomenon is anything mental that happens. Sense-perceptions are occurrent mental phenomena; and so are acts of ratiocination and emotional epiphanies. (This list is far from complete.)

But sometimes the word "thought" refers to the relatively stable *structures* that *underlie* mental occurrences. "For the first twenty years of my academic career," a philosopher once said to me, "I thought that Platonism was false." This person was describing a stable, structural constituent of his mind, not a mental *happening.*

If, by "thought," Hume means "mental happening," then, contrary to what

he alleges, there's nothing incoherent in the idea of a mind that, during some period, isn't populated by thoughts.

If by "thought," Hume means "stable mental structure," then Hume *is* right to say that there couldn't be thought-free minds. We saw why this is so in Section 6.1.6. Given that, if selves were distinct from their thoughts, there could be thought-free minds. So, given that there cannot be thought-free minds, selves are constituted by the thoughts that belong to them.

So, after a fashion, Hume is right. Just as he says, there is nothing to one's self other than the thoughts one has. But this claim of his isn't right in the way he thought it was. He thought that one's self consists of the thoughts (=fleeting mental occurrences) that occupy it, and not of the thoughts (=stable structures that create and regulate such occurrences). We've already discussed why Hume didn't even grant the existence of stable mental structures or, indeed, of structures of any kind.

7.0 Objectual identity revisited: Which causal sequences are thing-constitutive?

Earlier in this chapter, we saw that every object—every rock, tree, and vase—is a causal sequence. But not every causal sequence constitutes some one object. Some constitute more than one. Others constitute none.

Here's an example of one that constitutes no objects. Suppose that billiard ball #1 strikes hitherto stationary billiard ball #2, causing it to roll. Ball #2 then strikes hitherto stationary billiard ball #3, causing it to roll. And so on. This sequence of events doesn't constitute a single entity; this sequence isn't like a rock or a tree or an ice cube.

Here's an example of a causal sequence that constitutes more than one object. Let S_1 be the series of events constituting some seed *and* the process by which that seed turns into a tree and *the* tree itself. S_1 is a causal series—a textbook case. But it constitutes two distinct objects, not one. The seed isn't a mini-tree. It's a tree-precursor, not a tree, just as a gamete is a person-precursor, not an actual person. (Unlike people, gametes don't have thoughts or fall in love or play the harmonica.)

Two questions arise. What is the difference between those causal sequences that are thing-constitutive and those that are not? Second, what is the difference between thing-internal change, on the one hand, and wholesale transformation, on the other?

7.1 Some general features of thing-constitutive causal sequences?

Nothing that endures for more than an instant is *not* a causal process. But what is distinctive about those causal processes that we think of as objects (e.g., frozen ice cubes) is that the changes constitutive of them occur within certain narrow and fixed limits.

There's nothing unchanging about any ice cube—even one that for the last fifty years has been inside in an undisturbed deep freezer. Given such ice-cube, every particle constituting it is in context flux; and the ice-cube itself is nothing but a flurry of activity.

But even though those particles are all moving vigorously, they're moving only in ways that allow the flurry they compose to retain more or less the same properties. The *flurry* doesn't change, even though the things composing it do. The reason the flurry doesn't change is that, although its components are changing incessantly, the way in which they change doesn't change.

Thus, what we *usually* describe as "changes" are really second-order changes: changes in the changes themselves. Each molecule composing a given ice-cube is an arena within which ceaseless change occurs. But so long as the ice-cube remains intact, there are no significant changes in the relations borne to one another by these different arenas of change. That's why the ice-cube is unphased by such changes. But if the ice-cube is smashed or melts into a puddle, those mutual relations *are* significantly changed. That's why the ice-cube does not survive such changes.

7.2 Change vs. obliteration

Not all objects are as static as ice cubes, and our analysis must be refined if it is to deal with more dynamic entities (e.g., a living human body). Like a rock or an ice cube, a living body consists of individual particles each of which

bristles with activity. But where living bodies are concerned, there are changes, sometimes dramatic ones, in the relations that these arenas of change bear to one another. This happens whenever an organism develops in a positive direction, as when maturation or recovery strengthen it, or in a negative direction, as when old age or illness enfeeble it.

Thus, not all second-order changes jeopardize a thing's existence, and sometimes they're actually necessary for it. When do second-order changes destroy existence and when do they sustain it? "Such changes are existence-unfriendly," it might be suggested, "when they're abrupt or discontinuous and they're existence-friendly when they're not." But this isn't the right answer, since the melting of an ice cube, though decidedly thing-*un*friendly, is as smooth a process as the development of an organism, which is a decidedly thing-friendly process.

"Changes are thing-preserving," it will be said, "when they're *structure* preserving. The general morphology of an adult human being is similar to that of an infant. There are differences, of course, but they're internal to more fundamental similarities. Thus, the changes that connect the infant to the adult preserve structures. In general, thing-sustaining changes are structure-sustaining changes."

This position obviously has merit, but it won't quite do. There are cases where, without going out of existence, some one thing changes in fundamental structural respects. The butterfly doesn't resemble the caterpillar; but there is some *one* organism that is first a caterpillar and then a butterfly. So why is it that, whereas the ice cube goes out of existence when it melts, the caterpillar is transformed, not obliterated, when it turns into a butterfly?

7.3 Prediction- and explanation-facilitation—the essence of objecthood

Puddles by no means defy prediction. If you see a puddle that is expanding, you may know roughly how long it will continue to expand and roughly what its shape will be when it stops expanding. But in addition to being of limited scope and low quality, such predictions presuppose *ad hoc,* circumstance-specific knowledge. The predictive payoff is minimal, given the size of the

epistemic investment.

By learning more about the relevant variables (e.g., the shape of the floor, the rate at which the puddle is expanding, etc.), you can increase the predictive payoff. But given a vast improvement in one's knowledge of these variables, there would be only a slight improvement in the predictive payoff. To have any chance of predicting with any precision what the puddle's shape will be when it stops expanding, one would need extraordinarily high-resolution knowledge. Further, interference-effects would guarantee that, no matter how good your measuring instruments were or how adept you were at using them, your knowledge of initial conditions would not exceed a rather low threshold. And supposing that, within these limits, you acquired as much knowledge as possible of the relevant variables, any predictions you could make on the basis of that knowledge would be of minimal scope. Given high-resolution knowledge of the puddle's state at a given time, you could make low-probability, low-resolution predictions about the puddle's state at times soon thereafter.

But the story is very different where butterflies, human beings, and other *bona fide* things are involved. Given a knowledge of slight disturbances to a caterpillar's anatomy, one can make high-resolution, high-probability predictions about the anatomical condition of the butterfly into which, in due course, it will metamorphose. Easily acquired low-resolution knowledge of an animal's state at a given time may yield a high-resolution of its state at a much later time.

Given a pregnant mother who consumes large enough amounts of alcohol, it can be predicted that, 50 years hence, the person to whom she gives birth (supposing that person to be alive) will be severely damaged, and it can be said with some precision how that person will be damaged. The other, brighter side of the coin is that, given easily acquired, low-resolution knowledge to the effect that a certain person, while in his formative years, is given certain kinds of stimulation, allowed certain kinds of freedom, etc., one can make high-probability, high-resolution predictions about that person's condition in late adulthood.

Of course, butterflies and children aren't just things—they're organisms. And not everything that is true of organisms is true of all things. But points

relevantly similar to the ones just made hold of non-organisms. A sculptor who is chipping away at a hunk of marble knows that, barring some extraordinary event, the shape had by the hunk when he's done chipping away at it is the shape it will keep for centuries to come. The people who built the Washington Monument knew that, barring some extraordinary event, it would remain standing, its shape and general physical condition more or less unchanged, for hundreds if not thousands of years to come. In general, a limited knowledge of a thing's state at a given time yields high-probability, high-resolution knowledge of that thing's state at a much later time.

A knowledge of this principle underlies all physical labor. It underlies the manufacturing of cars, houses, fishing rods, and elevators. All of this activity would be pointless were it not for the prediction-friendliness inherent in thinghood.

These predictions are defeasible, of course.[299] Well-built houses are destroyed by earthquakes. Well-educated children from loving homes turn out to be moronic ghouls. But it's *despite,* not *by virtue of,* its being a thing that a sculpture or child or anything else has a counterpredictive future. Embittered by his lack of artistic ability, some unknown kook breaks into the Louvre and destroys famous sculpture X. The world is shocked. Nobody could have predicted that the world would ever be without this masterpiece. X's future was indeed an unpredictable one. But until its thing-status was completely undermined, X's future *was* a predictable one. Thus, X's future was unpredictable *only* to the extent that X's thing-status wasn't determinative of it. So given only that X's future was not predictable, it doesn't follow that things *qua* things aren't prediction-enablers; and given that X's future *was* predictable to the extent that its thing-status had a hand in it, it follows that things *qua* things *are* prediction-enablers.

In general, to be a thing *is* to be a prediction enabler. It would, of course, be viciously circular to define a "thing" as a prediction- and explanation-enabling *thing*. But when we replace the second occurrence of the word "thing" in the last sentence with "complex of events," the circularity goes away. Thus, to be a thing is to be a prediction-enabling complex of events (or "event-complex," as we'll say). By the same token, an event-complex that is

predictively inert is, to that extent, a non-thing.

Prediction and explanation are conjugates of each other. An explanation uses known laws to go from effect to cause. (I know that Smith is now belittling Brown. I know that, when belittled, insecure, touchy people, like Brown, become angry. So I predict that Brown will become angry.) The corresponding prediction uses the same laws to go from cause to effect. (I know that Brown is now angry. I know that, when belittled, insecure, touchy people, like Brown, become angry. I also know that, the circumstances being what they are, nothing else could have caused Brown to become angry. So I explain his current anger in terms of his being insulted a few moments ago by Smith.) Thus, a swarm of events is a thing to the extent that it is prediction- and explanation-conducive.

7.4 Thinghood is a matter of degree

The property of being prediction-enabling comes in degrees, as does the property of being explanation-enabling. A brick house is richer in these properties than a plywood house; a plywood house is richer in them than a cloud; a cloud is richer in them than a breeze.

A consequence of our analysis is that thinghood comes in degrees. Some things are "thingier" than others. I accept this consequence. A brick house is more of a thing than a cloud; it's more "thingy." This means only that it provides a broader and more stable basis than does a cloud for prediction and explanation. Since this is what we are inclined to think anyway, it isn't to the discredit of our analysis that it has this consequence.

We are inclined to think that things must literally be solid, like rocks and ice cubes, and unlike puddles. So far as there is a connection between actual solidity and thinghood, it is that, within the narrow horizons within which we live, solid things typically provide a firmer basis for prediction than non-solid things. But non-solids can be as thingy as rocks. Though they're gaseous, as opposed to solid, stars are things to no less a degree than rocks. The reason is that, even though they're not solid, they have the dynamic integrity of things that, in terrestrial contexts, are typically had by solid things.

7.5 The contextual nature of thinghood

Despite its internal lack of cohesiveness, a galaxy can be highly thing-like in the context of an explanation of why certain extremely large-scale astronomical events are occurring—of why, for example, there is a marked tendency for objects in a certain very large region of space to head in a certain direction. And despite its *high* degree of internal cohesiveness, a rock can be as unthing-like as, in most contexts, gusts of air would be. To creatures whose minds moved millions of times more slowly than ours, rocks would be as unthing-like as gusts of air are to us. And to creatures whose minds moved millions of times more quickly than ours, gusts of air would be as durable, cohesive, and predictable as marble statues are to us.

Thus, a given system of events can be a thing—in certain respects, at certain junctures—in some explanatory contexts, while being unthing-like—in different respects, at those same junctures—in others. But that doesn't mean that things are created by our explanations.

8.0 Time-travel an incoherent concept

It is debated whether time-travel is possible. Thanks to our analysis of what it is to persist in time, we can shed some light on this problem.

Setting aside the trivial and irrelevant fact that we travel forward in time at a rate of one minute/minute, the statement "time-travel is possible" has three distinct meanings:

(a) One can "leap" from one time to some non-adjacent time, i.e., one can move in a *discontinuous* manner from one time to another.

(b) One can travel forward in time at a rate of *more* than one minute/minute.

(c) One can travel *backwards* in time.

It is generally thought that (a)–(c) are empirical questions that fall within the jurisdiction of physics. In actuality, they are questions of logic, not of empirical fact. Moreover, each of (a)–(c) is logically false and therefore couldn't possibly be correct.

Let's discuss (a) first. In this chapter, we have seen that anything that persists for any amount of time is a ***causal series*** (a continuous series of

states of affairs such that, given any two of them, one is the (in)direct cause of the other).

An illustration may be helpful. At any given moment during its existence, your car consists of an incessant flurry of mass-energy displacements; and your car is thus a series of "event-flurries," as we'll call them. Twenty-four hours ago, your car was constituted by one event-flurry. Right now it's constituted by a different event-flurry. (Let S_1 be the first event-flurry and let S_2 be the second.) And, of course, my car is currently constituted by some event-flurry. (Let S_3 be that flurry.)

Bearing all of this in mind, imagine the following. While your car was parked in your driveway, somebody crashed into it, severely damaging it. Your car therefore bears little resemblance to what it was twenty-four hours ago. It follows that S_1 and S_2 are qualitatively very different from each other. By contrast, S_1 and S_3 are as similar to each other as their being in different places permits. This is sheer coincidence: S_1 and S_3 are, we will suppose, as causally disconnected from each other as any two non-simultaneous events on this planet can be. It follows that my car, as it is right now, is just like your car, as it was twenty-four hours ago, before it was damaged. It also follows that S_1 is much less similar to S_2 than it is to S_3.

Be all of this as it may, the car in your driveway is, whereas the car in my driveway is not, numerically identical with the car you owned twenty-four hours ago. Mediating between S_1 and S_2 is a continuous series of events such that, given any two of them, one is the (in)direct cause of the other. No such series mediates between S_1 and S_3. That's why (i) the car in my driveway isn't the very same car as the one you parked in your driveway twenty-four hours ago, their uncanny mutual resemblance notwithstanding. And that's also why (ii) the car sitting in your driveway is the very same car as the one you parked in your driveway yesterday, their *lack* of mutual resemblance notwithstanding. Obvious generalizations of these points establish our thesis that objects are identical with causal series.

A consequence of the fact that objects are such series is that a given object ceases to exist the moment the corresponding causal series is severed.

If, instead of just being damaged last night, your car was vaporized, my car, though an exact duplicate of your (pre-vaporized) car, still isn't numerically identical with it.

Given that an object ceases to exist the moment the corresponding causal series is severed, it follows that, so far as it isn't logically false, the statement:

(i) Smith disappears in 2009 and re-appears in the year 2023

collapses into the statement:

(ii) Smith is annihilated in 2009 and a Smith-*like* entity that isn't actually Smith *ex nihilo* comes into existence in the year 2023.

A question will help clarify the reasoning behind this last assertion: Where is Smith during that fourteen year interval? He's either nowhere or he's somewhere. Let's suppose that he's nowhere. In that case, for the reasons illustrated in connection with your car, whatever it is that pops into existence in the year 2023 isn't numerically identical with Smith. So let's suppose that he's somewhere—that, at any given instant during
that time period, he is in some place or other. But in that case he's no more time-traveling than anything else that exists during that interval. If he's next to some one tree during that entire period, he's no more time-traveling than that tree. If he's going from place to place, he's no more time-traveling than any one of the satellites that will orbit the earth during that period.[300]

So depending on whether he's somewhere or nowhere, Smith is either permanently annihilated in 2009 or he's no more time-traveling than anything else that exists during that period. Therefore, (a) is false.

To see why (b) is false, we need do little more than repeat some of the points just made. Suppose that Smith travels from 2009 to 2023 at a rate of one million minutes/minute. We've just seen that Smith cannot make this journey in a discontinuous manner; he cannot disappear from one time and reappear at some other, non-adjacent time. So Smith must make this journey in a continuous manner. So at any given moment during this interval, Smith must be in some place or other. It follows, as we saw a moment ago, that he is no more time-traveling than anything else that exists during that interval.

The supposition that Smith is traveling forward in time at a rate of one million minutes/minute thus turns out to collapse into the supposition that he isn't time-traveling at all. Given any positive number n, an analogous argument shows that continuous forward time-travel at a rate of n minutes/minute collapses into forward time-travel at a rate of one minute/minute.

There is another, quite different problem with the concept of discontinuous time-travel. Let W_1 be a universe where (i) is true and (ii) is false; and let W_2 be an otherwise indistinguishable universe where (ii) is true and (i) is false. What happens in the one universe at any given time coincides with what happens in the other universe at that time. It follows that, even if (*per impossibile*) an occupant of the one universe knew exactly what happened in that universe at any given time, he would not, in virtue of that fact, have any legitimate reason for preferring (i) to (ii). *A fortiori,* no possible observation that one might have would provide one with any reason for believing oneself to be in W_1 as opposed to W_2. Since observation couldn't possibly provide one with such a reason, only logic could do so. But logic cannot do so, since (i) is logically false, as we've already seen.[301]

Similar points hold in connection with discontinuous backwards time-travel. In the movie *Back to the Future,* young Marty McFly vanishes from the year 1985 and shows up in the year 1955. This means that, in the year 1985, the causal sequence that constitutes his existence is severed. Like everything else that persists for any length of time, Marty McFly is a causal sequence. It follows that, so far as it isn't incoherent, the statement

(1) in 1985, Marty McFly traveled backwards in time to the year 1955 and is otherwise unchanged

collapses into:

(2) Marty McFly was annihilated in 1985, and in 1955 there appears *ex nihilo* somebody who is numerically distinct from him but otherwise just like him.

Also, given obvious adaptations of what we said a moment ago, there is

no conceivable event that, were it known, would provide more support for (1) than for (2).

Since, therefore, there couldn't possibly be any possible observation-based reason for preferring (1) to (2), the only possible reasons for doing so would be of a logical nature. This means that there are *no* reasons for doing so, given that (1) is incoherent. The concept of discontinuous backwards time-travel is therefore a broken one.

What about the idea of *continuous* backwards time-travel? For argument's sake, let's suppose that, during his continuous journey from the year 1985 to the year 1955, Marty McFly is right next to some stop-sign that was built in 1920 and stays in the same place for several centuries. That means that, at any given moment during that thirty year period, Marty McFly is no more time-traveling than the stop-sign. Each is just sitting there.

There are, of course, some important differences between Mary McFly and the stop-sign. First of all, whereas the stop-sign is crushed or burnt or otherwise destroyed in some pedestrian way, Marty McFly must simply vanish in the year 1985. (If he doesn't, then he's still hanging around in the year 1985, in which case
he obviously hasn't gone back in time.) And, whereas the stop-sign's presence in the year 1955 is to be explained in the customary way, Marty McFly just popped up *ex nihilo.*

Bearing these points in mind, let W_3 be a universe in which

(3) Marty McFly continuously traveled backwards in time from the year 1985 to the year 1955 but is otherwise unchanged

is true, and in which

(2) Marty McFly was annihilated in 1985, and in 1955 there appears *ex nihilo* somebody who is numerically distinct from him but is otherwise just like him

is false; and let W_4 be a universe that is exactly like W_3 *except* that, in it, (2) is true and (3) is false. Echoing what we said earlier, there is no conceivable event that, were it known, would enable one to determine whether one was in

W_3 as opposed to W_3, or *vice versa.* Which, taken together with the fact that being annihilated in 1985 didn't exactly tighten young Marty's grip on existence, suggests that, other things being equal, (2) is the better way to model the observational data.

There is another, even deeper issue. For there to be backwards time-travel of any kind, there must be back retroactive causation: some events must be the effects of *later* events. Marty McFly's appearing in 1955 must be an effect of some event occurring in 1985. But there are empirical and also strictly logical reasons for thinking backwards time-travel an impossibility.

If Relativity Theory (RT) is right, which we'll henceforth assume that it is, the statement "E_2 is an effect of E_1, but E_1 post-dates E_2" is tautologically false. According to RT, order in time is *defined* in terms of causation: anything caused by a given event *ipso facto* follows that event. Thus, the view that one could travel backwards in time is in the same category as the view that triangles could have five sides.

A well known consequence of RT is that a given event may precede one event relative to one framework, be simultaneous with it relative to some second framework, and follow it relative to some third framework. Some conclude from this that backwards time-travel *is* possible.

This is not the right inference to draw. For reasons that aren't important here, the framework-relative nature of temporal order is a *consequence* of, among other things, the fact that any effect of an event *ipso facto* follows that event. In other words, it's a consequence of the fact that "E_2 precedes E_1 and is also an effect of it" is tautologically false.

Like any other scientific theory, RT is empirical. It's based on observation; and acceptance of it is warranted only as long as our observations are consistent with it. Nonetheless, the principle that causes *ipso facto* precede their effects is an ***analytic,*** not an ***empirical,*** truth. (We'll henceforth refer to this principle "CT"—short for "causality and temporal order.") So CT is true even if RT is false.

"But how can CT be analytic," it will be asked, "given that RT is an observation-based theory?" Two reasons. (1) Not every statement composing an empirical (observation-based) theory is *itself* observation-based. Every

theory comprises statements of pure logic (e.g., "no true statement entails the negation of anything entailed by any of its consequences") that aren't observation-based. (2) Not everything learned with the help of observation is learned *through* observation. If one learns some complex mathematical theorem by reading a proof of it, one learns that theorem *through* observations of ink-deposits; but it isn't *by observing* such deposits that one learns it. The observations of ink-deposits merely occasion or "trigger" the purely logical ratiocinative activity that *does* lead to a genuine knowledge of that principle.[302] Somebody with 20/20 vision who has carefully studied those ink-deposits (i.e., who has studied those ink-deposits themselves, as opposed to their meanings) may completely and utterly fail to understand the corresponding proof; whereas somebody whose vision is just good enough to see some of the key lines of the proof may, on the basis only of those observations along with the thoughts they precipitated, understand it extremely well.

Also, somebody who has heard the proof stated but hasn't read it may, on the basis only of those auditory observations along with the thoughts they precipitated, thoroughly understand the proof. Ink-deposits bear no physical resemblance to noises. Therefore, it isn't by studying ink-deposits that one learns the proof.[303] So given only that empirical discoveries led to our becoming aware of CT, it doesn't follow that CT was itself discovered empirically.

And, in fact, CT couldn't have been discovered empirically. Assuming it a given that it was Smith's pushing the button that caused the elevator to come, there is no conceivable event that, were it known, could warrant a revision one's subsequent judgment that the pushing of the button preceded the arrival of the elevator.

Einstein himself very clearly said that CT *was* not empirical hypothesis. He also said (unfortunately for me) that it was a convention—a *definition* of "order in time"—and, therefore, not a substantive truth. I would suggest that, although he described CT as a definition, he chose this particular definition, as opposed to others, because he saw in it a substantive truth whose roots lie in the very concept of what it is for one event to precede another.

Chapter 17

Causality

1.0 The basics

Things make things happen. Smith pulls the trigger, causing the gun to fire. Jones throws the vase against the wall, causing the vase to break. There are causes and there are effects.

Astonishingly, some philosophers have denied the existence of causation. Why? Because many philosophers are ***empiricists,*** and given an empiricist outlook, it is hard to accommodate the presumption that events make events happen. According to empiricism, all knowledge is strictly observation based. Thus, if you know it, one of two conditions is met: (a) you observed it (in other words, you saw it, heard it, felt it, etc.); or (b) you inferred it from what you observed, *provided* that you used no inference rules other than those of whose legitimacy direct observation apprised you. (See Chapter 13.) According to many empiricists, we cannot see or otherwise observe causal ties. We see events, but not forces. We see rocks strike windows, and we see windows shatter upon being thus struck. But we don't see anything that, given such a collision, *compels* the window in question to shatter. Since we don't observe causal connections, we don't have good reason to believe them to exist.

But this argument involves a major non-sequitur. We don't observe electrons or quarks, but we have good reason to believe them to exist. How are causal ties different? Given that it's methodologically permissible to posit unobservable viruses and particles, why isn't it equally permissible to posit unobservable causal ties?

"Because," we are told, "the concept of a causal tie, unlike that of a proton or a virus, is an incoherent one. For that reason, it wouldn't do any good to posit causal ties. Causal ties, if they exist, are occupants of space-time, like rocks and vases and chairs; and they are therefore just as incapable as chairs and rocks of being the "glue" that, supposedly, keeps the universe from powdering into so many renegade events." But notice that the second argument, unlike the first, isn't an empirical argument at all. It's a strictly

logical one. It is to the effect that forces are in the same category as square circles and therefore couldn't *possibly* exist and, therefore, couldn't possibly be observed. Thus, the first empirical argument against the existence of causal ties is parasitic on a non-empirical, purely ***a priori*** argument, and is therefore only as good as the latter. (An *a priori* argument is one that has its basis, not in observational data, but entirely in considerations of coherence and logic.) And the latter argument isn't a good one. The concept of *persistence* is itself a causal notion, and anything that persists is, for that very reason, a causal sequence. Thus, we observe causal sequences whenever we observe anything. And we *knowingly* observe them. That is, we don't observe them only in the

attenuated sense in which we are observing H_2O molecules when we look with the naked eye at bodies of water. This line of thought will be developed in Section 5.3.

Much of this chapter will be spent evaluating the contention that we don't see or otherwise observe causal ties. Although it has been enormously influential, it's of doubtful accuracy, it's of questionable coherence, as we've just seen. And the more it's scrutinized, the more incoherent it proves to be. For example, the very notion of spatial *occupancy* is a causal notion, since a given thing's occupying a given region R ipso facto alters the trajectories and velocities of other would-be and soon-to-be occupants of that region.

These points generalize. A "steel" object that lacked the requisite degree of malleability or thermal conductivity would ipso facto not be steel. Malleability and conductivity are causal properties. x is more malleable than y if, other things being equal, less force is needed to cause x's shape to change than is needed to cause y's shape to change. x has a higher degree of thermal conductivity than y if, other things being equal, less energy is needed to raise x's temperature by a given amount than is needed to raise y's temperature by that same amount. Thus, to describe a given object as being of made of steel is to say that, other things being equal, it will react to situations in ways in which it wouldn't otherwise react.

In describing something as "self-identical" or as "being such that 1 + 1 = 2," one is not making a causally loaded claim. But those characterizations say nothing about anything. In being told that x is "self-identical," one isn't being

told how x is different from, or similar to, water or steel or the planet Neptune or anything else. The same is true of all causally sterile characterizations. For that reason, anything that says anything about anything in the spatiotemporal world has causal content.

2.0 Causation a relation between events, not objects

But before we enter these stormy and controversial waters, let us make some non-controversial points about the logic of the concept of causality.

First of all, it is only *events* that can be causes. *Objects* are not causes. The rock is not a cause. The rock didn't cause the window to break. What did so was some event involving the rock (e.g., the rock's striking the window at a certain time).

Sometimes the way we speak suggests that, in our view, it is objects, not events, that are causes. For example, people say "Carter was the cause of the recession in the late '70s," "Hitler was the cause of World War II." But these statements are elliptical for "such and such events involving Carter (Hitler) were responsible for the recession in the late '70s (World War II).

In connection with this, we must distinguish between something's being a cause, on the one hand, and its having causal properties, on the other. Non-causes may have causal properties. The rock isn't a cause, as we discussed, but it has causal properties. The rock's having a certain mass is among its causal properties, meaning that its having that mass is, or can be, a (partial) cause. (Not every property of the rock is a cause. The rock is such that it either is or is not identical with the number. But the rock's having that property doesn't cause anything to occur.)

We must also distinguish conditions from events. A condition is a relatively persistent state (e.g., the room's having a certain temperature). An event is a *change* in conditions. But conditions *themselves* seem to be causes. For example, the water's having a certain temperature is what caused the spaghetti in it to soften. At the same time, conditions appear not to be events. An event is, by definition, a change, whereas a condition is, it would seem, the absence of change. So if conditions are non-events and are also causes, then, contrary to what we earlier said, it isn't events alone that are causes.

This argument overlooks the fact that conditions *consist* of events. The

water's having a given temperature *consists* in its containing particles that have certain kinetic properties. There is no heat where there are no events. A person's being healthy (i.e., being in a condition of health) consists in that person's body doing what it should. There is no life, and therefore no health, where there are no metabolic events. And although events do necessarily constitute changes, those changes may be, and often are, internal to conditions of changelessness. For the water's temperature to remain at 101° for 5 minutes is for the changes undergone by its constituent molecules to remain within certain limits; it is for their mean kinetic energy to stay within certain bounds. For the water's temperature to change is for their behavior to exceed those limits. What we call "conditions" are patterns of *constant* change (i.e., of change that occurs in a uniform manner). What we call "events" are *changes in the manner of change.* When the water's temperature increases, the *way* its constituent molecules move about changes. When the water's temperature stays the same, they still move about—they still undergo change—but the *way* they do so doesn't undergo change: those changes fall within certain limits. So conditions *can* be causes. But since conditions, far from being the absence of events, *consist* of them, this is in no way inconsistent with our contention that it is events, and events alone, that are causes.

Nothing non-spatiotemporal can be a cause. There are causes only where there are changes, and changes happen in time. So universals, if they exist, are causally impotent. Many have concluded that, for this reason, they cannot be known and, therefore, that there is no legitimate reason to grant their existence. The idea is that there must be some kind of a causal connection between two things if one of them is to know of the other. For me to see the rock, it is necessary that light rays bouncing off it affecting certain sensory surfaces of mine. For me to know of Socrates' existence, I must be on the receiving end of testimony, the vehicle for which is always some causal process (noises, light rays bouncing off of ink marks), that can be traced back to some state of affairs involving Socrates. One obvious way to generalize this line of reasoning is to say that there is *never* awareness where there isn't a causal nexus.

For reasons given in Chapters 8 and 9, this conception of conception is erroneous; under no circumstances does a causal connection between two

things *by itself* mediate the one's having an awareness of the other. There is always a non-causal component to awareness, even to sense-perception, which, of all the many forms of awareness, is the one in which a causal connection between subject and object is most obvious. Further, for reasons given in Chapter 2, we do have good reasons to countenance the existence of universals, and no good reasons not to.

3.0 Causes as INUS-conditions

The things we describe as "causes" are really *partial* causes. Supposing it correct, the sentence "Smith's drunkenness is what caused the accident" must really be elliptical for "Smith's drunkenness is a *part* of what caused the accident." Had he been driving a car with new tires on an open country road, instead of a car with bald tires on a busy city street, Smith might not have crashed *even* if he'd been drunker than he actually was. So it wasn't Smith's drunkenness or any other one thing that caused the crash, and was instead a confluence of circumstances. In some contexts—for example, those of a legal or ethical or psychological nature—the *relevant* partial cause might well be Smith's drunkenness; and that seems to be what we mean when we say that it was *the* cause.[304]

The *complete* cause of anything is very hard to identify. It is a matter of philosophical and scientific debate whether, given an event E happening at t, anything other than the entire state of the universe just prior to t can be the *complete* cause of E. So for all intents and purposes, the expression "*the* cause of occurrence E" never refers to a *sufficient* condition for E.

"Even though Smith's being drunk wasn't *sufficient* for his crashing the car," it will be thought, "it was *necessary* for it. Had he *not* been drunk, he wouldn't have crashed." This is false. Smith's being drunk wasn't necessary for his crashing. If he had been sober, but had been driving too fast on an icy road, he might well have crashed.

J.L. Mackie (1917–1981) made it clear what is meant by "the cause of E," as this expression is typically used. Smith's being drunk wasn't a necessary condition for his crashing.[305] He could have crashed if sober. But his being drunk was a necessary condition for his crashing *under the*

circumstances. Given that he wasn't driving a high-performance car on a country road during a clear day, and given that he *was* driving a low-performance car on a rainy night on a busy street, Smith's being drunk *was,* we may suppose, a necessary condition for his crashing. For it could be that, had he been sober, Smith would have narrowly avoided crashing, despite the rain, darkness, and faulty steering mechanism, and that it was only these things *plus* Smith's drunkenness that led to the crash. So, if it is to be correct, the statement "Smith's drunkenness was *the* cause of the crash" must be taken to mean Smith's being drunk was *under the operative circumstances* necessary for his crashing.

There are just a few more points to make. First, given that Smith did crash the car, it follows trivially that *a* sufficient condition for his doing so was fulfilled. Second, whatever that sufficient condition was, it wasn't *necessary.* Suppose that what led to the crash was Smith's driving a low-performance vehicle on a dark, rainy night, while drunk. Even though, by hypothesis, that confluence of circumstance was sufficient for his crashing, it wasn't *necessary.* Smith might have crashed even while, stone cold sober, he'd been driving a high-performance vehicle in broad daylight.

Thus, Smith's being drunk was an *insufficient* but *non-redundant* part of an *unnecessary* but *sufficient* condition for his crashing. Smith's being drunk was thus an "INUS-condition" for his crashing. For this reason, Mackie says that, if it is to express a true statement, "E was *the* cause of event E*" must be taken to mean that E is an INUS-condition for E*. In a word, "the cause" of an event is an INUS-condition for that event.

4.0 Aristotle on causality

Aristotle (384–322 B.C.) distinguished between four different kinds of causality: efficient, final, material, and formal.[306]

x is the efficient cause of y if x makes y happen.

x is the final cause of y if y is an action on the part of some creature and x is that creature's objective in performing y. Final causes are goals or objectives.

x is the material cause of y if x is the material that y is made of.

x is the formal cause of y if x is y's form. (To expedite discussion, we'll henceforth omit the obvious qualifications relating to the fact that Aristotle didn't use English words.) Sometimes Aristotle uses the word "form" to refer to shape. But sometimes he uses it to refer to what he elsewhere refers to as "essences." Those properties that are essential to a given thing jointly constitute its essence; and a property is essential to a given thing if that thing could not lose that property without *for that very reason* ceasing to exist. The italicized phrase is intended to indicate that the dependence-relation in question is not of a *causal* nature, and is instead of a *logical* or *constitutive* nature. In other words, that relation is comparable to the dependence that triangles have on being three sided, not to the dependence that mammals have on air.

Aristotle's belief that *things* have essences is an incoherent one, and it's a special case of an incoherence that pervades Aristotle's conception of causality. It is essential to a thing's *being triangle-shaped* that it have three sides. But a given triangle-shaped object's continued existence needn't depend on its continuing to have three sides. Things change shape. It is essential to a thing's *being water* that it consist of H_2O molecules. But a given puddle's continuing to exist needn't depend on its consisting of such molecules. Things undergo chemical changes without ceasing to exist.

Contrary to what Aristotle thought, dependence-relations don't hold between objects. They hold between *events* or between *statements,* depending on whether they're of a causal or a logical nature. To say that x's having three sides is essential to x's being a triangle is to say that "x is a triangle" entails that "x has three sides." To say that x's consisting of H_2O molecules is essential to its being water is to say that "x consists of water" and "x doesn't consist of H_2O molecules" are incompatible statements. To say that treating children with kindness is essential to their becoming sane adults is to say that treating a child unkindly prevents them from becoming sane adults.

So-called formal causes don't hold between events (or states of affairs) or statements; they hold between things. Thus, the concept of a formal cause is an incoherent one. So, for the same reason, is the concept of a material cause.

The concept of efficient causality clearly *is* a coherent one. There is such a thing as efficient causality. We'll see in the next section that it isn't the only

bona fide form of causality. But it's the most fundamental one.

What about (so-called) final causes? If, as Aristotle seems to think, final causes are goals per se, he's wrong to describe them as causes. Goals per se do nothing. Goals are abstract objects and therefore do no causal work. What *does* do causal work is a given person's *having* a given goal. My *having* the goal that

(*) I, JMK, get in shape

obviously does causal work; it causes me to jog around the block for hours on end. But (*) per se does nothing. Like all other goals, (*) is a proposition. My *having* that goal is my intending to make that proposition come true. But my *having* a goal is a garden-variety efficient cause. In general, goals per se, being propositions, aren't causes at all. A given person's (or creature's) having a given goal *is* a cause. But it's a garden-variety efficient cause. So in saying there are such things as "final" causes, Aristotle is either making the trivial point that a person's having a goal can have effects *or* he is making the deeply confused point that goals per se can have effects.

4.1 The importance of the expression "in virtue of" in causal discourse

The expression "in virtue of" has an important place in philosophical discourse. (I will use "in virtue of" and "by virtue of" interchangeably.) Without this expression, or some equivalent, it would be very hard to say anything meaningful about causal explanation—or about non-causal explanation, for that matter. The meaning of this term is to be understood in terms of two parallel points, namely:

(i) it is states of affairs (or events), not objects, that have effects;

and

(ii) it is propositions, not objects, that stand in dependence-relations.

For reasons that will become apparent, (i) and (ii) can be condensed into:

(iii) dependence-relations hold between propositions, as opposed to objects.

A story will help us explain what (i)–(iii) mean. Despite being intellectually mediocre, Smith never loses an argument. He wins arguments,

not by addressing points with counterpoints, but by bullying his opponent into submission. In this way, Smith recently won an argument with Jones, who is Smith's superior intellectually. Thus:

(A) even though Smith won his argument with Jones, he did so not *by virtue of* making good points, but *by virtue of* bullying Jones.

What (A) means is that the cause of Smith's winning was *his being a bully,* and not *his being smart.* (Cause: Smith's being a bully. Effect: Smith's winning. *Smith* wasn't the cause of anything.) Thus, the occurrence of "by virtue of" in (A) is a way of encrypting a statement about states of affairs—of making them seem about objects, when in fact they are about states of affairs.

Even if Smith *did* make good points in the course of the argument, it doesn't follow that he won *by virtue of* making good points. It could be that his making good points had nothing to do with his winning and that it was his bullyism that was responsible for it. In that case,

(A*) even though Smith won his argument with Jones, and even though he made good points in the course of that argument, Smith won *by virtue of* his being a bully, and not *by virtue of* his making good points.

Importantly, Smith could have won *both* by virtue of being a bully *and* by virtue of winning good points. In other words, his bullyism and his making good points could have been co-factors.

Let us now discuss (ii). Suppose that there are only three triangular objects in the world, x being one of them, and also that these three objects are the only green objects in existence. In that case, although "every triangle is green" is true, "x is green *by virtue of* its being triangle" is false. The right statement is:

(B) "x is a triangle by virtue of its having three co-planar sides, each of them straight, such that any two of them intersect and such that not all of three of them intersect."

What (B) means is:

(C) The proposition that x is a triangle entails, and is entailed by, the proposition that x has three co-planar sides, each of them straight, such that any two of them intersect and such that not all of three of them intersect.

In what follows, I will use the term "fact" to mean "spatiotemporal state of affairs." So a "fact," as I'll be using this word, is a distribution of mass-energy.[307]

Suppose that, for reasons of physical law, nothing could be triangular without being green, and nothing could be green without being triangular. In that case, given some green/triangular object x, the fact in virtue of which x is green—in other words, the distribution of mass-energy in virtue of which x was green—would be spatially and causally indistinguishable from the fact in virtue of which x was triangular. Nonetheless, it wouldn't be in virtue of its being a closed three-sided (etc.) figure that x was green—it would be in virtue of its having certain reflective properties. And it wouldn't be in virtue of its having those reflective properties that x was triangular—it would be in virtue of its being a three-sided, straight-edged (etc.) figure. This shows that the meaning of (B) is (C). When occurring in statements affirming logical (as opposed to causal) dependencies, "in virtue of" denotes a relationship holding, not between things *or* even facts, but between propositions.

Given that the occurrence in (A) of "by virtue of" does denote a relationship holding between states of affairs, what are we to say? One response, of course, is to say that it's ambiguous. But, in addition to appearing very wrong, this response is unnecessary. (A) doesn't express an *entailment.* But it does express *a* dependence-relation holding between statements. The truth of *Smith won the argument* is, in this context, dependent on the truth of *Smith bullied his opponent.* In general, "in virtue of," when occurring in causal contexts, can be interpreted as denoting a form of interpropositional dependence; and we may therefore condense (i) and (ii) into

(iii) Dependence-relations hold between propositions, as opposed to things.

4.2 Functionalism, content-externalism

According to a doctrine known as "functionalism," it is *wholly* in virtue of what its causes and effects are that a given brain-state (or physical entity of any kind) falls into a given mental category. Anything that has certain causes and certain effects is ipso facto a belief that snow is white or a desire for a

world peace or a vitriolic dislike of a certain piece of music.[308]

Functionalism has no plausibility when it comes to sensations or, more generally, anything that has ***phenomenal content.*** But it's widely taken to be correct when it comes to so-called "dispositional" mental entities such as beliefs and objectives. (I put the word "dispositional" in quotes because, for reasons that I'll now give, I reject this way of characterizing the mental entities in question.)

Thus restricted, functionalism may initially sound plausible. But it has some grisly problems that a story will help to expose. While standing right in front of you, Jones, a demented acquaintance of yours, puts what is clearly arsenic in a glass of water. He then tries to make you drink the water. Jones is very strong and is also a trained martial artist; so while holding the glass in one hand, without spilling any of it, he puts in you a joint-lock and tries to pour the contents of said glass down your throat. You shut your mouth, and you twist and squirm. Miraculously, you break free and run out the door.

What caused all of this behavior? A good part of it was your belief that Smith was trying to make you drink arsenic. Another part of it was your belief that, were Smith to succeed in making you drink arsenic, you'd suffer some adverse consequence. Yet another part was your wish to not die. And, of course, other ***propositional attitudes*** of yours were determinative of your behavior. Supposing that x is the part of your brain that comprises the various physical structures and events that mediate these different propositional attitudes, it isn't just *x* that makes you squirm and fight, and so on. Rather, it's *x's being a belief that Jones is about to make you drink arsenic.*

But functionalism cannot accommodate this obvious fact. According to functionalism, it is in virtue of what x causes and what causes it that x is a belief that Jones is about to do that. The causal relations fix the content, not the other way around. This means that, if functionalism is right, x's being a belief that Jones is about to make you drink arsenic does *nothing.* In general, given any entity O that mediates mental state M,

O's being an instance of M without effects. But this means that, if functionalism is correct, then the mental does nothing; it has no causal powers at all. Of course, functionalism is consistent with, and in fact depends on, the presumption that the brain-states that mediate mental states have

causal powers. But since functionalism is inconsistent with the fact that *in virtue of being mental states,* brain-states are powerless, functionalism amounts to the denial that the mental (*qua* mental) has any causal powers at all. Given the fact that anything spatiotemporal ipso facto has causal powers, the reason being that occupancy of a time or a place *is* a form of causality, it follows that functionalism denies the very existence of minds.

According to a doctrine known as "***content-externalism,***" two people, Smith and Twin-Smith, who are exactly the same, leaving aside the remote causes of their current conditions, can have thoughts with different contents. So, because Smith's current visual perception was caused by light that bounced of Mary, whereas Twin-Smith's current visual perception was caused by light that bounced off of Twin-Mary, otherwise identical Smith and Twin-Smith are now having thoughts with different truth-conditions and, therefore, different contents. (Smith is thinking *that person*—viz. Sally—*is tall.* Twin-Smith is thinking *that person*—viz. Twin-Sally—is tall.)

To appreciate just what the content-externalist says, we must distinguish causal from constitutive necessity. Obviously people have different thoughts. And obviously differences in people's personal pasts *create* many of the cognitive divergences between them. But the content-externalist is saying that, given two people who, like Smith and Twin-Smith, are *now* atom-for-atom duplicates of each other, and whose respective pasts haven't so much as made a dent in the one person that they haven't made in the other, those two people *will think different* things. They'll have thoughts with different contents. One of them will think about Sally (gold, water, etc.), while the other thinks about Twin-Sally (fool's Gold, etc.)

This view is incoherent. If Smith and Twin-Smith were to switch places, but everything else were held constant, neither would behave or think differently from how he would otherwise do so. Thus, *to the extent* that Smith's mental content is fixed by factors that aren't operative in Twin-Smith's case, Smith's thinking isn't at all different from Twin-Smith's. Any differences between Smith's and Twin-Smith's thoughts have some basis other than the differences in their distal causes. So to the extent that mental content is to be understood along content-externalist lines, it is without any causal powers at all, and therefore doesn't exist.

As for the content-externalist's contention that Smith is thinking that *Sally* is tall, whereas Twin-Smith is thinking *Twin*-Sally is tall, that is a non-starter. What Smith's eyes tell him isn't that *Sally* is tall. They tell him that *a* person who has certain features is tall. Given only what his eyes are telling him, Smith doesn't know whether he's seeing Sally or some Sally look-alike. To be sure, he may well know that he's looking at Sally. But that knowledge won't be based strictly on the content of visual perception. It will be based also on background knowledge that he brings to the present context.

Content-externalism embodies a failure to appreciate the concept meant by the phrase "in virtue of." True-Smith and Twin-Smith aren't doing the same things. Smith is one place; Twin-Smith is in another. Smith is looking at one person; Twin-Smith is looking at anther. But it isn't *in virtue of* any differences in their respective mental content that Smith and Twin-Smith diverge in these respects. Since, therefore, no differences between Smith and Twin-Smith, or by obvious extensions of this argument, between any two people, are to be understood along content-externalist lines, mental content is causally important, and therefore non-existent, to the extent content-externalism is the right theory about it.

Tyler Burge (1946–), a leading exponent of content-externalism, deals with this by arguing that the notion of causality should be revised, so as to accommodate content-externalism. Were two qualitatively identical objects x and y to switch places but everything were held constant, this switch would be experimentally undetectable. But Burge says that x and y nonetheless have *different* causal properties. By thus reanalyzing the concept of causality, he safeguards content-externalism from the objection to it just considered.

This position borders on incoherence. It's brazenly ad hoc. And it's unnecessary, since the data is easily modeled without it. (See Chapters 6, 8, 9, and 18.) A thorough discussion of content-externalism is found on pages 207–507 of my book (2007) *Conceptual Atomism and the Computational Theory of Mind.*

4.3 Program-causality[309]

We obviously know of many cases of one event's causing another. We are inclined to think that, in each such case, the one event is the *efficient* cause of

the other. Working jointly, Frank Jackson (1943–) and Philip Pettit (1945–) have shown this to be false. Given almost any known instance of one event's causing another to occur, the one event is the *program-cause,* not the *efficient* cause, of the other.

What does the term "program cause" mean? Some scenarios will help us answer this question. There is a kettle on the stove. You turn on the gas. The resulting flame flames underneath the kettle to heat up and thus causes the water to boil. Let F be the flame. We're inclined to think that F caused the kettle to heat up.

(i) F (or, in any case, F's having a certain temperature) caused the kettle to heat up.

But only a tiny part of F's surface was in contact with the underside of the kettle. Had the events at those contact points been present, the effect on the kettle would have been the same.

Another example: statements such as:

(ii) high-inflation led to crime, which, coupled with xenophobia on the part of the populace, gave the government a powerful incentive to scapegoat a foreign nation and go to war with them.

But causes are much more specific than (ii) suggests. *Crime* doesn't do anything, and neither does *xenophobia.* Things are caused by specific people, performing specific criminal or xenophobic acts.

Last example:

(iii) The impact of the rock on the window caused the window to shatter.

This case is similar to the first. Only a small part of the surface of the rock came into contact with the window—and, in fact, those parts of the rock's surface came into contact with specific parts of one of the window's two surfaces. Had the events at those contact points happened, the effect on the window would have been the same, even if no rock had struck the window.

But if we say that (i)–(iii) are just wrong, our grounds being that a "cause" of an event E can only be an event that is directly responsible for E's occurrence, then just about every single causal statement anyone makes will be wrong. But statements like (i)–(iii) obviously convey valuable

information. So rather than say that they're all just wrong, we should look for an interpretation of the word "cause" that ***validates*** those statements.

Jackson and Petit have done just that. The form of causality mentioned in (i)–(iii) is *program-causality.* x is a program-cause of y if x consists of events that, in their turn, are directly responsible for y. A country's being poor consists of various specific states of affairs—various specific people being without funds in various specific situations. The specific situations that *constitute* a condition of wide-spread poverty will directly cause crimes. And so on. F heats up the pan in the sense that among the events constitutive of F are those that heat up the pan; and it is in the same sense *mutatis mutandis* that the impact of the rock shattered the window. Thus, Jackson and Pettit have successfully validated ordinarily causal statements.

They make other important points. One would think that high-resolution causal information is better than low-resolution causal information. But this isn't always so. What caused the house to burn down?

(SA)[310] Arson on Smith's part was the cause. He threw several Molotov cocktails at the house and up in flames it went.

SA obviously doesn't give us the micro-story. For argument's sake suppose that some long sentence MS does give us the micro-story. MS will be worthless in terms of giving us information about Smith's character, and it will be equally worthless in other, comparables respects. There are many principled regularities of which this particular incident is an instance, and we blind ourselves to all but a few of them if we reject any story that

isn't the micro-story. So the macro-story—or, rather, the various different macro-stories—aren't inferior versions of the micro-story; they aren't precursors to it that we'll chuck when we have it. They expose regularities, structures, and counterfactual truths omitted by the micro-story. This shows that biology isn't just bad physics, that psychology isn't just bad physiology—it shows that different disciplines are dealing with different program-causes, and therefore distinct and very possibly *sui generis* nomic connections—that happen at least sometimes to be instantiated by the same efficient (micro) causes.

Predictions are almost always made on the basis of a knowledge of

program-causes. To predict what's going to happen, one must know of *structures* that can be depended on to yield certain outputs, given certain inputs. So far as one lacks such knowledge, one must make predictions on an ad hoc basis, and there are two reasons why this is a problem. First, sub-structural facts tend to be small—literally small. (Something is a structure only to the extent that it's a cohesive aggregate of events. Something fails to be a structure only to the extent that it's an isolated event. Thus, lone events tend to be smaller than structures by orders of magnitude.[311]) And when they can be known, the process of learning about them tends to disrupt them in prediction-foiling ways. (See Chapter 14.) Also, structures comprise back-up mechanisms that ensure a certain outcome even if, because of some hidden variable, some rogue, structure-unfriendly event makes it through the gauntlet. If a salesperson working for megacorporation X unexpectedly quits, another takes his place. If one police officer doesn't report for duty, another steps into the breach. It's of the essence of structures that a certan amount of causal redundancy be incorporated into them. Where there isn't such redundancy, predictions are easily foiled, since there is nothing to absorb the ripple-effects created by some rogue occurrence.

Given the dependence of predictive knowledge on structures and, therefore, on program-causes, it follows that knowledge of explanations is comparably dependent. The reason is that explanation is reverse-prediction. One predicts X when one identifies current conditions that, given known mechanisms, make X's occurrence likely. One explains X when one identifies past conditions that, given known mechanisms, resulted in X's occurrence. The concept of a program-cause, which was, so far as I know, first clearly analyzed by Pettit and Jackson, is therefore central to the concepts of prediction and explanation and, thus, to the theory of knowledge.

5.0 Hume's first analysis of causality: causality = regularity

According to David Hume, nothing makes anything happen. There are no forces or structures underlying the various events that occur. There are just the events.

Why does Hume think this? Because he accepts a very strict form of

empiricism, this being the doctrine that what is known is either (i) directly observed to be the case or (ii) is inferred from what is directly observed in accordance with inference-rules of whose legitimacy direct observation has apprised one.

Do we observe connections between events, Hume asks? No—we only observe the events. Thus, Hume believes, so far as you have empirical grounds for saying that the collision of billiard balls B_1 and B_2 caused previously stationary B_2 to start moving, it's because you know the following three conditions to be fulfilled (let E_1 be the collision and let E_2 be the subsequent displacement of B_2): (1) E_1 immediately preceded E_2; (2) E_1 and E_2 are adjacent; and (3) any event similar to E_1 immediately precedes, and is adjacent with, some event similar to E_3.

Conditions (1) and (2) explained: If there were a spatial or temporal gap between the two events, we'd be unjustified in believing the first to be a *direct* cause of the second. Hence conditions (1) and (2). Condition (2) explained: If events similar to E_1 weren't typically adjacent with, and immediately followed by, events similar to E_2, we'd have no reason to think that E_1's occurring had anything to do with E_2's occurring.

Hume thus proposes the following analysis of causation:

(HC_1[312]) E_1 causes E_2 to occur if E_1 immediately precedes and is adjacent with E_2, and any event similar to E_1 immediately precedes and is adjacent with some event similar to E_2.[313]

5.1 Problem #1 with Hume's analysis: It doesn't distinguish between actual causation and pseudo-causation

This analysis is untenable. A story will help us identify one of the reasons why. Your dog Fido is an infallible earthquake detector. Whenever an earthquake is about to strike, he senses it, and barks hysterically and runs around in circles. One day, Fido starts barking hysterically and running

around in circles. Sure enough, an earthquake strikes right then and there.

Hume's analysis entails that Fido's antics *causes* the earthquake to strike. Fido's antics immediately precede the earthquake. (So condition (1) is satisfied.) Fido's antics are adjacent with it. (So condition (2) is satisfied.) And such antics on Fido's part are *always* adjacent to an immediately ensuing earthquake. (So condition (3) is satisfied.) Given that Fido's antics obviously didn't cause the earthquake, Hume's analysis is wrong.

The real cause of the earthquake presumably created some disturbance that Fido detected, that being why he engaged in his antics. Thus, some one thing was responsible both for Fido's antics and for the earthquake. The two effects fulfilled the conditions described by Hume. For this reason, the sequence they formed mimicked a bona fide causal connection.

When some one thing E_1 has two effects E_2 and E_3 that *appear* to be related as cause and effect, respectively, we say that the sequence consisting of E_2 and E_3 is a "pseudo-causal process" or, alternately, a case of "epiphenomenal causation." Hume's analysis cannot distinguish epiphenomenal causation from actual causation.

5.2 Problem #2: Hume's analysis entails that events can't be space-time neighbors without one of them causing the other to occur

If E_1 and E_2 are any two spatiotemporally adjacent events, it's easy to find categories K_1 and K_2 such that E_1 and E_2 belong to K_1 and K_2, respectively, and such that any instance of K_1 immediately precedes, and is adjacent with, some instance of K_2. Though he doesn't know it, Bob is due for a heart attack. Right before the heart attack, Sally gently hugs him. Given the almost certainly correct supposition that nobody who is *exactly* Sally's height has ever hugged anyone who is *exactly* Bob's height, it follows that any hug involving people with heights m (Sally's height) and n (Bob's height) immediately precedes and is adjacent with a case of somebody of height n having a heart attack. (So if K_1 is the category of hugs involving people with heights m and n, and K_2 is the category of people with heights n having heart

attacks, any event E_1 that belongs to K_1 immediately precedes and is adjacent with some event E_2 that is a member of K_2.) So Hume's analysis has the false consequence that the hug caused the heart attack. Given any two spatiotemporally adjacent events x and y, a similar argument shows that, if Hume's analysis is right, x caused y. Since not all such pairs of events are related as cause and effect, Hume's analysis is false.

In response to this it will be said that:

> The categories in question must be relevant ones. K_1 and K_2 won't do. They're the wrong ones. All you've shown is that Hume's analysis must be modified. Instead of:
>
> (HC1[314]) E_1 causes E_2 to occur if E_1 immediately precedes and is adjacent with E_2, and any event similar to E_1 immediately precedes and is adjacent with some event similar to E_2,
>
> Hume should have said:
>
> (HCI#) E_1 causes E_2 to occur if E_1 immediately precedes and is adjacent with E_2, and any event *relevantly* similar to E_1 immediately precedes and is adjacent with some event *relevantly* similar to E_2.

This proposal doesn't work. Given any two spatiotemporally adjacent events, each of those two events will belong to many different categories. Hume's analysis is powerless to tell us whether the one caused the other *until* we select the relevant categories. *But we can't select the relevant categories until we have the very knowledge that Hume's analysis is supposed to give us.* It is only because we already know that the hug isn't what caused the heart attack that we know that K_1 and K_2 aren't relevant categories.

A related point is that HC1# is viciously circular. Supposing that E_1 directly causes E_2, for K_1 and K_2 to be the "relevant" categories is for it to be the case that it's *because* E_1 and E_2 belong to them that those events are related as cause and effect. Thus, HCI# collapses into the viciously circular statement that:

(HC1+) E_1 causes E_2 to occur if E_1 immediately precedes and is adjacent with E_2, and anything like E_1 causes something like E_2 to occur.

5.3 Problem #3: Hume's analysis is viciously circular

According to Hume, x is a cause of y only if x and y are space-time neighbors. Thus, Hume analyzes the concept of causality in terms of the concepts of space and time. Were it to turn out that the concepts of space and time are *themselves* to be understood in terms of the concept of causality, Hume's analyses would be viciously circular. He'd be analyzing causality and in terms of itself and thus not analyzing it at all.

In fact, the concepts of space and time *are* to be understood in terms of the concept of causality, and Hume's analysis *is* viciously circular for this reason. According to modern physics, E_1 precedes E_2 iff it's possible for a signal beginning with E_1 to be sent to E_2. E_1 and E_2 are simultaneous iff it isn't possible for a signal beginning with either to be sent to the other. Order in time is thus to be understood in causal terms.[315]

So is order in space. Events happen in different places if, and only if, something can affect the one that cannot affect the other. If x and y are in coincident places, nothing can affect the one without affecting the other. If x and y are not in coincident places, something can affect the one without affecting the other. Thus, x and y are in different places exactly if something affecting the one may fail to affect the other. So if one says that x and y occupy distinct but adjacent places, or that they occupy some one place, one is *ipso facto* making a causal statement. Thus, Hume's analysis of causality is circular, given that it analyzes ‹x causes y to occur› in terms of ‹x is spatiotemporally adjacent with y.›

Here is a similar line of thought. It would be meaningless to say that object x moved but that its causal liaisons to other object didn't change for that very reason. If I approach you, my causal relations to you change, and so do your causal relations to me. But after those changes are set aside, there is nothing left. There is no "pure" change in position underlying all the causal

changes.[316]

Also, the concept of spatial occupancy is a causal concept. What is it for me to be occupying this place? It is for it to be harder than it would otherwise be for you to enter this place. It makes no sense to say: "x is occupying spatial region R, but x's occupancy of R has no effect on what things besides x must do to occupy R." To occupy a place *is* to make it harder, other things equal, than it would otherwise be for other things to enter that place.

Hume analyzes causality in terms of the concept of spatial occupancy. E_1 causes E_2 only if they occupy adjacent, or coincident, spaces. But since E_1's (or E_2's) occupancy of a place is *itself* to be understood in causal terms, Hume's analysis analyzes causality in terms of itself and therefore doesn't analyze it at all.

A corollary of the fact that spatiotemporal relations are causal relations is that, contrary to what Hume alleges, we *do* see causal processes and, moreover, we see them *as* causal processes. We see nothing instantaneous. Everything we see persists, if only for a short while. Anything that persists *is,* for that very reason, a causal process; and anything that is seen *as* persisting, which is everything that is seen, is, for that very reason, seen *as* a causal process.

There are two views as to what space and time are. According to the one view, space and time are not empty containers. Spatiotemporal positions are relations between events, and there thus couldn't be a space-time empty of events. If this view is right, it's meaningless to say some object *just* approaches some other. For x to precede y isn't for x's *position* to bear a certain relation to y's position. Given *any* spatiotemporal relation R, for x to bear R to y is not for x's position to bear R to y's position. There are no event-independent positions. Relations between positions are mediated between relations between events, not *vice versa.* It thus
makes no sense to suppose that there could be bare spatiotemporal interrelations or, therefore, bare changes in such interrelations.[317]

According to the other view, space and time exist independently of events. A consequence is that spatiotemporal positions exist independently of the events that might (or might not) occupy them. Spatiotemporal relations

holding among events are to be understood in terms of spatiotemporal relations holding among positions in space-time. This means that there are absolute positions in space-time. If this view is right, it *does* make sense to say that one event *just* precedes some other event, or that one event *just* happens to be equidistant between two other events, and so on.[318]

Bearing these points in mind, consider the statement that:

(IP[319]) x immediately precedes and is adjacent with y,

and suppose IP to be true. If the first view is the right one, there are no event-independent positions. In that case, x's being adjacent with y is a relation that holds *directly* between x and y, and to say that some position bears a certain relation to another position is to say that some *event* bears a certain relation to some other event.

If the second view is the right one, the fact described by IP consists in x's occupying some position that bears a certain relation to a position that is, in turn, occupied by y. In other words, it is only because the *positions* they occupy are adjacent that x and y are adjacent. Thus, the second view presupposes that there are event-independent, and thus absolute, spatiotemporal positions.

Since Hume's analysis presupposes the correctness of the second view, Hume's analysis is refuted by the very empirical and philosophical arguments that show the concept of absolute position to be an untenable one.

5.4 Hume's denial of causality based on a pseudo-empirical, crypto-a priori argument

Hume says that we don't observe forces. But this seems false. When you are trying to lift a heavy suitcase or trying to arm wrestle a much stronger person, you are, so it would seem, observationally aware of a force.

Hume considers this very point. In response to it, he says that what you are aware of is a mere sequence of events. You have various perceptions. Some are visual—you see your opponent's sweaty, determined face. Some are tactile—you feel your opponent's increasingly sweaty hand. Some are kinesthetic—you feel your arm moving the wrong way. But, Hume says, you

don't perceive any force in all this—even though, because you are in the grips of some wrong theories, you think you do.

But this seems very wrong. It's a datum that you *feel* your arm being forced to move in a way you don't want it to be. That doesn't meant that this feeling is accurate. Maybe it's an hallucination, like one's visual perception of a pink elephant. But it's a datum that you *have* such a feeling, even though it isn't a datum that it's accurate. Plus, this response of Hume's is decidedly contrived. If, in describing your arm-wrestling experience, you left out the part about how you *felt* as though your arm was being moved the wrong way, you'd be seriously underdescribing your experience.

Thus, *if* the observational evidence is taken at face value, there are forces. This doesn't mean that the observational evidence *should* be taken at face value. But it means that Hume's denial of causality cannot, if legitimate, be based on strictly empirical grounds. And this means that his empirical argument against causation is only as good as some other, purely a priori argument.

That other argument, it will be recalled, is as follows. (What follows is not a quotation.) "The very concept of a force is an incoherent one. Were they to exist, forces would just be more constituents of the physical world. They would thus be no more capable than vases and rocks and so forth of constituting the 'cement' of the universe. The idea of a force that *wasn't* that sort of cement—that *didn't* glue events together—is a contradiction in terms. Therefore the concept of force is an incoherent one." We saw that vases and rocks *are* such cement. Any given physical object is a series of events any two of which are "cemented" together. Therefore, this argument of Hume's is no good. Therefore Hume's analysis of causality is no good.

6.0 Hume's second analysis of causality: the counterfactual analysis

Hume gives two analyses of causation. We've just discussed the first. According to the second, for E_1 to cause E_2 to occur is for it to be the case that E_2 wouldn't have occurred unless E_1 had occurred. This is known as the ***counterfactual analysis of causality*** (CF).

The motivation for CF is clear. The elevator comes because somebody pushed the button. If the button hadn't been pushed, the elevator wouldn't have come. The gun went off because somebody pulled the trigger. If the trigger hadn't been pulled, the gun wouldn't have gone off. No cause, no effect. Thus, ‹E_1 caused E_2 to occur› seems to be equivalent with if ‹E_1 hadn't happened, E_2 wouldn't have happened.›

CF must be distinguished from the related, but weaker claim that:

(CF*) Given any two events E_1 and E_2, such that E_1 precedes E_2, E_1 caused E_2 to occur just in case E_2 wouldn't have occurred unless E_1 had occurred.

What is the difference between CF and CF*? For argument's sake, let's suppose that, given any mathematical truth T, Smith knows T, and that given any *non*-mathematical truth T*, Smith does *not* know T*. In that case,

(SM*) T is a mathematical truth iff Smith knows T.

But it *isn't* true that:

(SM) *For* T to be a mathematical truth is for Smith to know T.

Similarly, given only that CF* is true, it doesn't follow that CF is true. But if CF is true, it does follow that CF* is true.

CF and CF*, though different statements, are *extensionally* equivalent. In other words, given any two events x and y, if CF entails that x caused y, so does CF*, and *vice versa.* If CF is to be correct, it must be extensionally correct—that is, it must entail that x caused y just in case x did cause y. Given that CF and CF* are extensionally equivalent, the one is extensionally correct iff the other is as well.

6.1 Why, even if it's extensionally correct, CF is false

Jones drinks poison and dies (at time t) as a result. (From now on, we'll leave the "at t" implicit.) It may well be true that:

(SB) Jones wouldn't have died if he hadn't drunk poison.

SB obviously entails:

(WB) In W, Smith didn't die,

where W is any possible world in which Smith didn't drink the poison but is otherwise just like ours (or, in any case, is as much like ours as Smith's not drinking poison allows it to be).

Question: Why doesn't Smith die in W? Answer: because nothing causes him to. No cause, no effect. In our world, the cause is present. In W, it isn't. WB's truth is to be understood *in terms of* the fact that, in our world, Smith's drinking the poison is what killed him. So SB no more constitutes an analysis of the statement "Jones died because he drank the poison" than "Smith believes that 3 + 3 = 6" constitutes an analysis of any fact about arithmetic. And just as Smith's believing that 3 + 3 = 6 is to be understood *in terms of* the truth of
that statement, so WB is to be understood *in terms of* the truth of SB and thus *in terms of* the fact that Smith's drinking poison is what killed him. In general, "E_2 wouldn't have occurred where E1 did not," when true, is to be understood *in terms of* the fact that E_1's occurring is what caused E2. So even if CF is extensionally correct, it's no analysis of the concept of causality.

6.1.2 The vicious circularity of CF, and an outline of a positive analysis of counterfactual truth

CF assumes that the concept of counterfactual truth is not itself to be explained in terms of the concept of causality. But that may well be false. The statement:

(SB) If Smith hadn't fallen off the building, he wouldn't have died

seems to be equivalent, or nearly so, with:

(SB*) Falling off the building caused an otherwise unthreatened Smith to die.

Not *all* counterfactuals register causal facts. For example:

(SH) If 6-feet tall Smith were half his current height, he'd be 3-feet tall,

is equivalent, or nearly so, with

(SH*) Smith is 6-feet tall entails anything that is 3-feet tall is half Smith's height.

Notice that SH is analytically true, whereas SB is non-analytically true. It thus seems to me that counterfactuals are just disguised causal or logical assertions and that, consequently, alternative worlds have no place in any decent analysis of them. In any case, these cases provide us with evidence that counterfactual truth is itself to be understood in terms of causality and, therefore, that CF is guilty of vicious circularity.

6.2 Why CF isn't extensionally correct

CF assumes that that there's no causal redundancy in the world—that, were a given mechanism to fail, nothing would step into the breech. But sometimes there are back-up mechanisms. If, for some strange reason, electrocuting condemned criminal Smith didn't kill him, the state would have found some other way to do so. Thus, the statement "the cause of Smith's death was electrocution" isn't equivalent with "if Smith hadn't been electrocuted, he would not have died." And this seems to show that CF is wrong.

Admittedly, CF has little trouble dealing with this case, and with many others of its ilk. Smith didn't just die; he died in a very specific way. And he wouldn't have died that way if, instead of being electrocuted, he'd been shot. This suggests that it's only when events are seriously underdescribed that there appear to be counterexamples to CF.

But this maneuver doesn't always work. A and B are both trying to push the same doorbell button. Since neither is aware of the other's presence, each acts exactly as he would have in the other man's absence. The button only has room for one finger. Both fingers get there at the same time, but A's finger, being the bigger one, crowds out B's. The requisite electrical signal is sent at time t. Had A been absent, B's would have struck the same doorbell; and, given the rapidity with which B's arm was moving, the electrical signal would have been sent at precisely t.

It may be that, were it not for A's presence, B wouldn't have pushed the doorbell in exactly the same *way* as A. (Our story may even require this.) But the signal is sent at the right time. And that's all that matters. It doesn't matter how exactly B pushed the doorbell. Doorbells are binary systems. The signal is either sent or it isn't.

"But maybe the laws of physics don't validate your point," one might

protest. “Maybe it’s a matter of physical law that nothing happens in *quite* the same way in a world where its actual cause is absent and some standin has to take its place.” (Let’s refer to this point as “LF”—short for “laws of physics.”) If correct, LF is a complete disaster for CF. First of all, LF is an empirical point. Empirical investigation might or might not bear it out. But even if it did bear it out, the mere fact that empirical research was needed to verify LF is enough to show that LF has no place in a philosophical analysis of counterfactuals. Such an analysis is supposed to identify what, as a matter of logic, is meant by statements like SB. It’s obviously an interesting question what the correct *physical* analysis of SB is. But here we’re doing logical, not physical, analysis; we’re making it clear what SB *means.* And surely what SB *means* doesn’t include some as-of-yet undiscovered laws of physics.

We may conclude that the counterfactual analysis of causation is a failure.

6.3 David Lewis’ analysis of counterfactual truth and his concomitant defense of the counterfactual analysis

Should it turn out that the concept of counterfactual truth has no logical integrity, it would follow that CF had no logical integrity. But we can’t know whether that concept has any integrity until we have an answer to the question: what does it mean to say that, if P had been the case, Q would have been the case?

The most famous analysis of counterfactual truth is due to David Lewis (1941–2001).[320] His analysis is beautifully clear and straightforward. According to Lewis, the meaning of:

(KN[321]) “if kangaroos didn’t have tails, they would fall over”

is this: “given any world W where kangaroos don’t have tails, kangaroos fall over, provided that W is as much like our world as is permitted by the fact that, in it, kangaroos don’t have tails.”

In general, the right analysis of counterfactual truth, according to Lewis, is this:

(LA[322]) "given any world W where P is the case, Q is the case, provided that W is as much like our world as is permitted by the fact that, in it, P is the case."

6.3.1 The role of other worlds in Lewis' analysis

Lewis' analysis obviously assumes that other "worlds" exist. This raises the question: what does Lewis mean by the term "world"? There are two options. Option 1: Like our world, these other worlds are space-time manifolds that are populated by objects and events. Option 2: These other "worlds" are representations of some kind or other (sets of statements, presumably). Let's consider each option.

For argument's sake, assume that these worlds are sets of statements. In that case, if LA is correct, counterfactuals merely register interpropositional relations. In other words, KN merely registers the fact that the *statement* "kangaroos don't have tails" bears some sort of relation (presumably, one of confirmation) to the *statement* "kangaroos fall over." But if *that's* the right analysis, then until this relation is identified, LA *fails* to say what counterfactuals are. And once it does identify that relation, it won't be necessary to talk about other worlds at all.

Lewis sees this and, for this very reason, he holds that the worlds in question are real worlds—space-time manifolds populated by events and objects. Lewis is thus a *realist* about other worlds.

6.3.2 Some consequences of modal realism

As you might predict, the standard objection to Lewis' analysis is that:

> We don't know if there are other worlds. What we *do* know is that, even if there are other worlds, we don't know anything about them. So nothing happening in these other worlds, if they even exist, justifies these beliefs or, therefore, provides the basis for our counterfactual knowledge.

That may be true. But let's start out by *granting* Lewis' assumption that counterfactuals, when true, hold in virtue of states of affairs that obtain in veritable alternative worlds. If we show that *given this assumption,* LA fails, then we've shown that it fails relative to its own terms and that there is thus no hope for it.

Let W be *our* world—the real world. In W, Smith intentionally shoots Jones, killing him. There were no other threats to Jones' life; therefore:

(SK) If Smith had not shot Jones, Jones would be alive.

If LA is right, SK means the following:

(LASK): Let W* be a world where Smith doesn't shoot Jones but that is as much like our world as that fact lets it be: in W*, Jones is alive.

In W, Smith doesn't *just* shoot Jones. That event has a cause. (It actually has a plurality of causes. But for the sake of simplicity, we'll just speak of "the cause" of that event.) Let E be that cause. Presumably E has as at least one of its ingredients Smith's intention to kill Jones. But whatever E is, it can't occur in W*, at least not in its entirety. So, in W*, either Smith *doesn't* intend to kill Jones and therefore doesn't shoot Jones; or he does shoot Jones, but Jones is wearing a bulletproof vest. It could be anything. To fix our ideas, let's suppose that, in W* Smith's hand seizes up just before he pulls the trigger; the gun drops; Jones gets away. Let E* be the event that, in W*, makes Smith's hand seize up.

What we said about E is true of E*. E* didn't just come out of nowhere; it itself had a cause. Let E^*_2 be that cause. And E^*_2 had a cause—and E^*_2's cause had a cause, and so on. So, in W*, we have a long chain of events (E^*_1....E^*_n) not a single one of which occurs in W. Depending on how far back this chain has to go, it could make W* *very* different from W. In fact, W* might be *so* different that it might not be able to contain a Smith or Jones to begin with; and even if it did, there's no guarantee that Smith would shoot Jones or that, if he did, Jones wouldn't be wearing a bulletproof vest, etc.

Events have causal roots. Two worlds cannot differ with respect to a single event unless they differ with respect to all of the causal antecedents of those events. But, given that those go back to the beginning of time, that means that they'll differ a lot—so much the one world becomes unable to

validate the counterfactuals of the other.

It may be that *some* events are completely and entirely uncaused. But, in W, that happens only where micro-events are concerned. Smith's shooting Jones (in W) is an extremely complex event, involving innumerably causally cohesive parts and having deep roots in the past. So even if *some* events in our world are uncaused, Smith's shooting Jones isn't one of them.

And even if it *is* one of them, it's useless as far as LA is concerned. A logical analysis of LASK can't assume the results of empirical facts not in evidence and, therefore, cannot assume that Smith's shooting Jones is totally uncaused. In fact, such an analysis, being of a logical as opposed to an empirical kind, can't make *any* empirical assumptions.

6.3.3 Some consequences of modal realism

But even if it could assume this, LA would still be in trouble. First of all, we can't say that, in W*, Smith's *not* shooting Jones is just as uncaused as uncaused as the event his doing so in W. In other words, we can't say that,

in W*, Smith's hand *just* seized up—that, in other words, there was no causal basis for it. For, to the extent that this event was *un*caused, any other event could have happened instead of it. We could, of course, just stipulate that, in W*, Smith's hand seizes up. But that won't help. If, in W*, Smith's hand *just* seizes up, then there will be at least one *other* world—call it W#—such that W# is just like W* except that, in W#, Smith's hand *didn't* seize up or, if it did, something else happened that led to Jones being killed. In which case, LASK will be false. For, contrary to what LASK says, there *would* be a world satisfying the relevant conditions in which Jones dies. It's not a good idea, in general, to posit causal slack in the worlds we want to validate our counterfactuals, since that same policy makes those very worlds unreliable in the way of validating our counterfactuals.

Lewis' own way of dealing with the problem is different from those just considered. In his view, some *miracle* happens in W* that prevents E from occurring there. So even though the natural laws in W* demand that E happen there (or, at any rate, demand it no less than the laws of our world demand it here), a "miracle" suspends these laws, and Jones lives.

Lewis is saying that, if a world validates LASK, there is some miracle in it (at a certain time, in a certain place). So W*, by virtue of being able to validate LASK, *also* validates the following counterfactual:

(MLASK) If Smith had not shot Jones, then a miracle would have happened.

W* is, by hypothesis, a world where Smith doesn't shoot Jones but is otherwise maximally like our world. In W*, there's a miracle (at a certain time and a certain place). So W* makes MLASK true if it makes LASK true. But MLASK isn't a true counterfactual. And, of course, if a miracle is needed to validate MLASK, then a miracle will be needed to validate *any* counterfactual. What we just said about MLASK is true of "if Jones hadn't driven drunk, he wouldn't have crashed his car," "if Brown hadn't been so diligent in college, he wouldn't have gotten into a reputable graduate program," and so on. So assuming these counterfactuals were true, the same would hold of (e.g.): "if Brown hadn't been so diligent, a miracle would have happened," which is obviously a preposterous counterfactual.

Also, Smith's (actually) shooting Jones is an enormously complex event, involving trillions upon trillions of mass-energy displacements. In W*, those events can't happen. Which means that, in W*, it isn't just one miracle that has to happen: it's trillions upon trillions of them. Which, in its turn, makes W*'s relevance in this context extremely questionable. We're finding that, in order to get W* to validate LASK, we have to doctor it—to mutilate it. But the very fact that we have to tamper with it—that we must employ such artifices to find a world that, in keeping with LA, can validate LASK—practically proves that LA is wrong. A theory that only doctored data supports is a bad theory, and Lewis' analysis of counterfactuals is untenable or, at the absolute least, grossly undersupported.

6.4 How Lewis' realism about possible worlds almost saves the day for the Humean

If we're right, counterfactuals are identical with causal statements. According to David Hume, nothing makes anything happen. If Hume is right, it's hard to see how anything could, in any real sense, cause anything to happen. But this seems a little strange.

David Lewis accepts Hume's analysis. This is because, like Hume, Lewis is a hardcore empiricist and, for the same reasons as Hume, rejects the idea that there are forces. But Lewis rightly thinks that denying the existence of causality is going too far. He sees that, given a realistic attitude towards other worlds along with an acceptance of the counterfactual analysis of causality, the presumption that events cause other events to occur can be reconciled with Humeanism. (By "Humeanism" I mean the view that nothing makes anything happen.) For these reasons, Lewis says that:

(LC[323]) In our world, event x causes y to occur if, given any world W where x doesn't happen but is otherwise as much like our world as that fact permits, y doesn't occur.

But we've already seen why this stratagem fails. Given that x occurs in our world, a world W in which x doesn't occur will bear so little resemblance to our world that it won't be able to validate the target-counterfactual (viz. "if x hadn't happened, y wouldn't have happened").

6.5 The basic problem with Lewis' analysis (LA)

We obviously can't see or otherwise sense-perceive what is going on in other worlds; nor can we can sense-perceive evidence of what is going on in such worlds. Any knowledge that we have of what is going on in them is parasitic on our knowledge of counterfactual truths. So even if

(i) had x not happened, y wouldn't have either

is true just in case

(ii) given an x-free world W that is otherwise just like ours, y doesn't happen in W,

our *knowledge* of (i) has nothing to do with other worlds. This shows that we know (i) to be true because we know of some fact in *our* world that coincides with the fact registered by (i). Which, in its turn, shows that other worlds have nothing to do with counterfactual truth. Which shows that LA is false.

7.0 Humeanism and probability statements

In keeping with common sense, I believe that statements about the past confirm statements about the future and, more generally, that statements about the known confirm statements about the unknown.

It is hard to reconcile this belief with Humeanism. David Lewis is the only Humean who has even come close to succeeding in showing Humeanism to be compatible with the presumption that statements about the known can confirm statements about the known. To understand his argument, we must make a few preliminary points about probability.

Let's suppose that:

(BU[324]) At time t, there are 10 balls in the urn. Nine are white. The other one is black. And, at t, there's nothing in the urn besides these 10 balls. You know all this. But this is *all* you know that's relevant to the probability of your pulling a white as opposed to black ball out of the urn.

Even though BU isn't analytic, and neither are the statements composing it, it *does* seem to be analytic that ("BB" stands for "black ball"):

(BB[325]) given a single chance to pull a single ball out of the urn, the probability that you'll remove a black ball from the urn is 1 in 10.

That said, it seems obvious that, *given* BU, it is analytic that BB holds.

But it isn't analytic; it isn't even true. Given *only* BU, you are no more entitled to think that you'll withdraw a white ball than you are entitled to think that you'll withdraw a black ball or a green ball or an iguana.

It may be that, given BU *along* with some reasonable assumptions (e.g., that it isn't a law of physics that white objects cannot be displaced or that, at time t, all white balls in urns turn into black balls or green balls or iguanas), BB is analytic.

But if any one of the parenthetical assumptions is false, you have *no* chance, given BU, of pulling a white ball out of the urn. Since these assumptions, though perhaps obvious, aren't analytic, BB doesn't analytically follow from BU. For reasons analogous to those just given:

(WB[326]) Given BU, it's analytic that if only one ball is *selected* (not physically removed from the urn—just selected) there's a 90% chance that

it's white

is false if the word "selected" refers to anything that takes any time at all (e.g., a psychological act or a physical process). It is only if that word is taken in some totally atemporal sense that WB comes out true. But, thus taken, WB says only that nine out of 10 balls in the urn are white if nine out of them are white.

In general, when words like "probable" (and "probability," etc.) refer to *statistical* probability, they must be taken in a strictly non-temporal sense. The statement:

(*) given that, at time t, there are 10 apples in the barrel and only three oranges (and no other objects), your chances at time t* of selecting an orange are 3 in 13

is false. What is true is:

(*1) given that, at time t, there are 10 apples in the barrel and only three oranges (and no other objects), you're chances at time t of selecting an orange are 3 in 13,

but only if the term "selecting" is used metaphorically. More precisely, (*1) is true *only* if it coincides with the trivial claim that, if 3 out of the 13 objects in the barrel is an orange, the ratio at that time of oranges to non-oranges is 3 to 13.

Of course, even a Humean can grant the truth of trivial probability statements, such as (*). So if, by some remote chance, it could be shown that non-trivial probability statements (e.g., given that, in the past, all metal has expanded when heated, it will probably continue to do so) could be identified with trivial ones, the Humean *would* be able to accommodate the presumption, so crucial to thought and scientific explanation, that non-trivial probability statements are correct. David Lewis makes a noble effort to do just this.

7.1 Why, for this reason, Humeanism can't explain explanation

Let's start by describing the viewpoint of common sense. In saying that

Smith will *probably* disgrace himself at his job interview, given that he drank an entire bottle of gin right before, you are saying that he all but guaranteed an unacceptable performance at the interview. In saying that Brown's victory is probable, given that he drugged his opponent's water, you're saying that he all but necessitated his opponent's defeat. In these contexts, the word "probably" doesn't have the emptily statistical meaning borne by its homonym in statements like:

(NC) Given that better than nine out of 10 people don't own a pogo stick, a person chosen at random probably won't own one.

If common sense is right, statistical probability isn't explanatory probability (even though statistical probability is sometimes a consequence of explanatory probability: other things being equal, explanatory probabilities are expressed by statistical probabilities; and it's from this fact that the Humean conception of probability derives what little credibility it has).

The Humean denies all this, since he thinks that nothing makes anything happen. But this means that he has trouble accommodating the fact that anything can ever be explained. A world where nothing makes anything happen is one where nothing has any roots—where, no matter what the conditions are any outcome is just as likely as any other.

The Humean has three choices: (i) reject the presumption that some things can be explained; (ii) stop being a Humean, since Humeanism is inconsistent with this presumption; or (iii) try to reconcile Humeanism with this presumption. Obviously Humeans choose (iii).

They defend this choice by saying that, contrary to what we said a moment ago, the concepts expressed by words like "cause" and "probable" *are* of a strictly statistical kind. "Given E_1, E_2 is probable" means "most events of the same kind as E_1 are accompanied by events of the same kind as E_2." And, as previously discussed, "E_1 causes E_2 to occur" means "given any event of the same kind as E_1, there occurs an adjacent, subsequent event the same kind as E_2."

The latter claim is a complete failure. As we saw in Section 5.2, there is no way to say what it is for events to be "of the same kind" as E_1 or E_2 that

doesn't render Hume's theory of causality viciously circular. The very same argument *mutatis mutandis* holds in connection with the Humean theory of probability. But Lewis chooses to argue for this theory; and he does so very cleverly. So let's hear him out.

7.1.1 Lewis' argument[327]

We've just seen that, even within a Humean framework, probability judgments are rational when they merely redescribe statistics—that, for example, (#) may be rationally accepted if it's just a way of saying that, at a given time, one person out of two weighs more than 50 pounds. Probability judgments are irrational, given a Humean framework, when they involve extrapolations. Lewis' metaphysical apparatus gives him a way of sidestepping this problem.

Bearing all of this in mind, let's suppose that, up until time t, whenever a rock with a mass of m or more, moving with a velocity of v or more, that has collided with a window of thickness t or less, the window shattered upon impact. In that case, presumably, it's reasonable to judge that, the next time such a rock strikes such a window, the window will shatter. The non-Lewisian Humean is stuck; he has no way to validate this presumption. The Lewisian isn't stuck. Here's what he says[328]:

> (LA): Let C_1 be the class of worlds where, before time t, such windows shatter when struck by such rocks and where, after time t, such windows *continued* to shatter when struck by such rocks; and let C_2 be the class of worlds where, before time t, such windows shatter when struck by such rocks and where, after time t, such windows *didn't* shatter when struck by such rocks. C_1 has more members than C_2.

For the non-Lewisian Humean, judgments about what is likely to happen *next time* are great leaps into the unknown. For the Lewisian Humean, such judgments merely redescribe what is already there. To say that *next* time such a collision occurs the window is likely to break is to say that there are more worlds where it *does* break than where it doesn't. If it's *given* that exactly one

of the cards in the deck is a King of Spades, it follows trivially that, given an arbitrary card, it has a 1 in 52 chance of being a King of Spades. Our judgments about what will happen to windows in *future* collisions are equally trivial.

7.1.2 Why this analysis fails

LA is obviously a clever analysis. But it's false, and it's clear why. Any knowledge we have of what happens in other worlds after t will be at least as uncertain as our knowledge of what happens in our own world after t. There are two reasons for this. Any knowledge that we might have about anything in other worlds will be at least as uncertain as any knowledge that we have about anything in our own world. Moreover, any knowledge that we have of what happens after t in other worlds will be based on our knowledge, so far as we have any, of what happens after t in our world. So LA goes through only if it's assumed, question-beggingly, that we know what will happen in our own world's future.

But there are deeper problems with LA. Here we must remember what Hume said about inductive inference:

> (HA)[329]: Given only that such and such *has* happened, it can't be assumed that such and such will *continue* to happen—*unless* it's assumed, question-beggingly, that what *has* happened will *continue* to happen.
>
> In general, given only that such and such is true of the known, it can't be assumed that it also holds of the unknown—*unless* it's assumed, question-beggingly, that the unknown resembles the known.

An analogue of what HA shows concerning inductive knowledge holds with respect to knowledge of other worlds. Given only that, in *other worlds,* such and such happens after time t, it can't be assumed that such and such happens after t in *this* world—*unless* it's assumed, question-beggingly, that what holds in circumstances already known to us also holds in circumstances not yet known to us. So LA goes through only if it's *assumed,* question-beggingly, that inductive inference works.

One might object that

You (JMK) are missing the very point of LA. Supposing that there are 10 balls in the urn, and only one of them is white, the statement ("WB" is short for "white ball"):

(WB) the probability of an arbitrary one of them being white is 1 in 10

holds as a matter of definition and, therefore, can be affirmed without begging the question against HA. The beauty of LA is that it assimilates apparently non-innocuous probability judgments into the class of trivially true, totally innocuous probability judgments, such as WB. LA shows that:

(UB) the probability that future unsupported objects will fall is extremely high, given that all such objects (apart from helium balloons, etc.) have fallen so far,

which is non-trivial and (in this context) objectionable, is in the same category as WB, which is trivial and unobjectionable.

This point exposes a confusion that has vitiated the efforts of epistemologists and philosophers of science for centuries. This confusion is of such importance that it deserves a section unto itself.

7.2 Statistical vs. explanatory probability and the solution to the lottery paradox

The second word in WB is nothing more than a homonym of the second word in UB. In WB, the word "probability" has a purely statistical meaning. In UB, it has an *explanatory* meaning. Philosophers have tried to assimilate explanatory probability to statistical probability. And, while this may initially seem like good sense, it's as misguided as trying to show that a bank (financial institution) is a certain kind of bank (river's edge). This is made clear by the following paradox, commonly referred to as the "lottery paradox" ("LP," for short):

> (LP) There's a lottery. One million tickets are sold. There's one winner. So given any one ticket, the probability of its losing are 999,999 in 1,000,000. You buy a ticket. Let x be this ticket. You know that ("LT" is short for "losing

ticket"):

(LT) the probability of ticket x's losing are 999,999 in 1,000,000.

But do you *know* that it'll lose? No. It may even have been rational for you to buy the ticket. Given what you stand to lose by buying the ticket and given what you stand to gain in the remote chance that you win, it may have been in your interest to buy it, even after you take into account that the chances of your losing is 999,999 in 1,000,000. But this decision couldn't possibly have been rational if you *knew* that your ticket would lose.

In general, there's no knowledge where there isn't certainty. If, given the information at your disposal, there is *any* chance, no matter how small, that P isn't the case, then you don't *know* P. We all know (or so let us suppose) that:

(UB) the probability that future unsupported objects will fall is extremely high, given that such objects have fallen so far.

Is there *any* chance that, all of a sudden, unsupported objects will no longer fall? Yes. (The statement "all of a sudden, unsupported objects will no longer fall" isn't ruled out by the laws of logic. So it could, theoretically, be true.) Therefore, you don't *know* that they won't fall; and, by obvious extensions of this arguments, you don't *know* anything about the future (apart from sterile trivialities; e.g., "whatever happens, happens"). The very concept of "probable" knowledge is incoherent. The merely probable is ipso facto not known. To know it to be 99.9999% likely that unsupported objects will continue to fall is to *not* know that they'll fall.

The problem with LP is that it uses the word "probability" in an equivocal manner. All LT says is that there are one million tickets but only one winner. UB doesn't say anything of the sort. UB doesn't make a *statistical* point. It doesn't say: "nine times out of ten, unsupported objects fall." It doesn't say anything at all like that. UB makes a point about the relative merits of two competing hypotheses, these being:

(H_1) unsupported objects will continue to fall

and

(H_2) unsupported objects will not continue to fall.

UB is true if, given the data that we have, ^{H_1}'s being true would undermine the integrity of our belief system to a lesser degree than ^{H_2}'s being true. Given the information that we have right now ("SB" is short for "stilts beanie"):

(SB) there is only an extremely low probability that walking on stilts for an hour while wearing a beanie with a propeller on it is a foolproof cure for cancer.

Does SB say that better than nine cases of cancer out of 10 *aren't* cured in this way? No. It doesn't make *any* sort of statistical claim. SB is a claim, not about statistics, but about explanatory integrity. It says that more damage will be done to our belief system if it turns out that cancer *can* be cured in that way than if it turns out that it cannot. If, by some remote chance, cancer *can* be cured in that way, everything we "know" about medicine and pathology is false. But the damage wouldn't end there. Our medical beliefs are rooted in general assumptions about the causal structure of the world. (For example, given the information that Smith has gone into anaphylactic shock whenever he's eaten peanuts or anything with peanuts in it, even cutting-edge medical researchers will assume that Smith will go into anaphylactic shock if he eats peanuts that have been painted green. This assumption is rooted in generally held causal beliefs. No inside information is involved.) Were it found out that SB was false, many of the commonsense-based underpinnings of our medical beliefs would have to be reconsidered. But all of those beliefs remain intact if cancer cannot be cured in that way. It is improbable that walking on stilts (etc.) cures cancer to the extent that our current beliefs depend on its *not* being a cure. And it's probable that walking on stilts (etc.) isn't a cancer cure to the extent our beliefs depend on it.

In general, P is probable to the extent that our beliefs depend on it, and improbable to the extent that they depend on its negation.

7.2.1 Two kinds of probability

Explanatory and statistical probability are different things. Counter*statistical* events are innocuous, but counter-*explanatory* events are not. Smith rolls double-sixes five times in a row. Statistically unlikely—yes. Important—no. Must we change our belief system because of it? No. Need we reconsider our assumptions about the basic structure of the world? No. And it makes no difference whether it's five double-sixes in a row or a billion. The improbable sometimes happens, and that's the end of it. When it happens, we have a nice anecdote for cocktail parties; we don't have a situation worthy of scientific scrutiny. But we *do* have such a situation if it turns out that Smith is using telekinetic powers to make the dice come up sixes. For, in *that* case, some deeply held convictions of ours have to be revised.

In describing Smith's rolling double-sixes five times in a row as "improbable," what one is saying is very different from what one would be saying in describing Smith's having telekinetic powers as "improbable." In the first case, one is affirming a mere statistical improbability. One is saying only that such things don't *often* happen. In the second case, one isn't saying anything of the sort. One is saying Smith's having such powers is inconsistent with some fundamental beliefs of ours—with beliefs that, if jettisoned, would take a lot of other, obviously (or at least presumably) correct beliefs with them.

Given *any* event, there is a way of describing it that makes it improbable in the *statistical* sense. In fact, *any* given event, described at all precisely, has a statistical probability of vanishingly close to zero. I flip a coin. Nothing improbable about this. But it's a safe bet that no coin has ever moved in *exactly* the way this one did after I flipped it. Described as a case of my flipping a coin, the event isn't improbable at all. But that same event is extremely improbable if described as a case of an object's moving in manner M, where M is the exact way that it did move. But even though nothing apart from that coin has ever moved in manner M, that coin's doing so isn't a matter that deserves the slightest scientific scrutiny. Its moving in that way is consistent with everything we know about the world, and there is therefore nothing to be explained. But *there is* something to be explained if Smith turns out to have telekinetic powers, since his having such powers is improbable in the *explanatory* sense. It's irrelevant that it is also improbable in the statistical sense. *Every* event is improbable in the statistical sense.

My winning the lottery is improbable in the statistical sense, but not in the *explanatory* sense. And that's why, before the winner is announced, we can't say that we *know* it won't win: my winning is consistent with everything we know about the world. My sprouting wings in the next 5 minutes is improbable in the *explanatory* sense. And that's why we *can* say that we *know* it won't happen. It's *in*consistent with everything (or much, at any rate) that we know about the world.

7.3 Grue: the problem of counterinduction

A puzzle devised by Nelson Goodman[330] will clarify, as well as support, what we've said ("GA" is short for "Goodman's argument"):

> (GA): Let's invent a new word: "grue," to be defined as follows. Something is grue if it is green before Jan. 1, 2010, or blue thereafter. So a shirt that's green on Oct. 12, 2007, is grue, as is a shirt that's blue on Nov. 11, 2012. Let's assume that, before Jan. 1. 2010, all emeralds have been green. (From now on, we'll abbreviate "Jan. 1, 2010," as "2010.") In that case, all emeralds have been grue prior to 2010. Anything that is green-and-examined-before-2010 is also grue. So the fact that all emeralds examined before 2010 are green means that all emeralds examined before 2010 are grue. We are assuming for argument's sake that, given the fact that all emeralds examined before 2010 were green, we are inductively warranted in believing that all emeralds examined after 2010 are green. By that logic, it follows that, given the fact that all emeralds examined before 2010 were grue, we are inductively warranted in believing that all emeralds examined after 2010 are grue. But anything examined after 2010 that is grue is *blue,* not *green.* So, given that all emeralds before 2010 were green, our inductive grounds for believing they'll be *blue* after 2010 are as strong as our grounds that they'll be green.

The degree to which a given fact is *statistically* improbable depends on how it is described. Given some emerald x that was examined before 2010 and found to be green, "x is grue after 2010" is, from a strictly *statistical* viewpoint, equiprobable with "x is green after 2010" and also with "x is blue

after 2010." But that's no cause for alarm. Statistical improbability is innocuous. It's explanatory improbability that you have to worry about. But the explanatory probability of "x is green after 2010" dwarfs that of "x is grue after 2010" and, therefore, "x is blue after 2010." The supposition that objects don't change color (except under conditions that emeralds don't satisfy, at least not in virtue of making it to 2010) is deeply embedded in scientific theory and also in the commonsense underpinning of scientific theory. If it's wrong, then science is in serious trouble, and so is human knowledge in general.

In response to this, it will be said that:

> If GA is right, emeralds that "turn blue" at the beginning of 2010 *don't* change color. They stay the *same color*—grue. And emeralds that "stay green" *do* change color, since they go from being grue to having some other color. (Or, to be more precise, emeralds *do* change color at the beginning of 2010—but only *relative* to what one set of color-terms, these being green, red, etc. But emeralds stay the same color *relative* to different color terms—grue, bleen, etc.)

This isn't true. Pictures of green emeralds (taken before 2010) show them *not* to have the same color as blue things after 2010. Grueness and greenness are different *properties.* But they're not different *colors*. Grueness isn't a color at all—it's a color-*related* property. Let's say that x is a "squiangle" (rhymes with "triangle") if it's square-shaped before 2010 and triangle-shaped thereafter. If, for some reason, some object that was square before 2010 suddenly becomes a triangle, it's obviously changed shape. It used to have four sides. Now it has three. And a "squiangular" object that is square-shaped *after* 2010 has started—in other words, an object that had four straight sides, etc., before 2010, and continued to have that shape—hasn't *changed* shape; it hasn't gone from being a "squiangle" to being a square. It's stayed the same shape. The same goes for "grue."

7.4 Scientific explanation: Hempel's theory

Let us now turn our attention to Carl Hempel's important theory as to the nature of scientific explanation.[331] Hempel's theory is important because it

is the clearest, purest expression of the Humean conviction that all probability is statistical probability and also because of the consequent fact that, given the enormously powerful hold of Humean ideas on methodological thought, all theories of explanation can be understood in terms of Hempel's theory—either as versions of it or as reactions *to* it. Hempel's theory is known as the "deductive nomological" account of explanation:

> (DN) To explain an event E is to show that its occurrence is a logical consequence of the statement formed by conjoining statements describing known natural laws with statements describing the conditions under which E occurred.[332]

I drop an otherwise unsupported object. The laws of physics being what they are, the object must fall, given the circumstances. And to explain that event is to identify known natural laws that, when conjoined with a correct description of the circumstances, entail that the object will fall.

7.4.1 Hempel's analysis evaluated

What is meant by the term "law"? The most obvious answer is that a law is what is described by *any* true universal generalization—that is, by any statement of the form "all phi's are psi's." (As we'll see in a moment, this is not the answer that Hempel gives. But the problems with the answer that Hempel does give are to be understood in terms of the problems with this answer.) If the word "law" is taken in this way, DN has the consequence that the following counts as an "explanation" of the fact that the pencil on my desk is yellow ("NE" is short for "non-explanation"):

> Premise #1 (description of the relevant natural law): Everything is such that it's either yellow or it isn't a pencil on JMK's desk.
>
> Premise #2 (description of initial conditions): x is a pencil on JMK's desk.
>
> Conclusion (thing being explained): x is yellow.

But NE isn't an explanation of the fact that my pencil is yellow. The problem, of course, is that[333]:

(YP) everything is such that it's either yellow or it isn't a pencil on JMK's desk

doesn't express a natural law. But why not?

Because, says Hempel, it *doesn't support counterfactuals.* The statement:

(ME) metal expands when heated

expresses a law of nature because a piece of metal that isn't being heated *would* expand *if* it were heated. Nothing comparable holds of YP. Given an arbitrary pencil, YP doesn't entail that it *would* be yellow *if* it were on my desk. In general, says Hempel, laws are expressed by universal generalizations that support counterfactuals.

Hempel is surely right that a statement must support counterfactuals if it is to express a natural law. But DN collapses if laws are understood in this way. Let m be some unheated metal object. Why is it that m *would* expand *if* it were heated? Because, the laws of nature being what they are, heating m would *cause* it to expand. So the fact that m *would* expand if heated is to be understood *in terms of* the fact that ME expresses a natural law.

Remember what we said earlier about counterfactuals: they are thinly veiled causal statements. "Smith would be alive if he hadn't jumped" means: "jumping is what caused the death of an otherwise unthreatened Smith." So far as that causal claim is warranted, it is because there is known to be a law that, given the circumstances, demanded Smith's death. The counterfactual claim is to be understood in terms of a causal claim, and the causal claim is to be understood in terms of a statement describing a natural law. DN thus becomes viciously circular if "law" is defined as "what is expressed by a counterfactual-supporting universal generalization."

Causal claims are *themselves* explanations. In being told that lung cancer is what led to Smith's death, I'm being given an explanation, albeit an incomplete one, as to why he died. Counterfactuals, as we know, are identical with causal claims. It is only where explanations are *already* available that counterfactual claims are warranted.

By the same token, if we don't know that it was arsenic consumption that killed Smith—or, if not Smith, then somebody else—we have no basis for affirming the counterfactual that Jones, who isn't consuming arsenic, *would* die *were* he to do so.

Thus, we are warranted in affirming counterfactuals only where we *already* have explanations. So to understand the concept of law in terms of the concept of counterfactual truth *is* to understand the law in terms of the concept of explanation. Thus, if laws are understood in this way, DN is analyzing the concept of explanation in terms of itself.

So long as the concept of law is understood in terms of the concept of explanation, DN is viciously circular. Is there some other way to understand the concept of law? This question exposes a serious problem with DN. Hume tried to show that laws are *mere* concomitances. He denied the existence of causal ties; so he couldn't take laws to be descriptions of such ties. And he therefore couldn't take "metal expands when heated" to describe the *connection* that we presume to hold between x's being heated metal and x's expanding. He had only one option: the law isn't some structure or metaphysical glue *underlying* the fact that the property of being heated metal is always found in conjunction with the property of expanding; the law *is* that regularity. Laws *are* the regularities that non-Humeans typically regard as their *expressions.*

The advocate of DN is forced to take this position. So long as he thinks of laws as being given by statements that describe *structures* of some kind—the structures that are manifested by regularities—the advocate of DN is analyzing the explanation in terms of itself. For it isn't *until* there are explanations that there can be any good reason to posit such structures. The structure is posited to validate preexisting explanations. Which makes it viciously circular to understand laws in terms of causality. Which, in turn, makes it viciously circular to understand laws in terms of counterfactual truth, given that counterfactual truths are to be understood in terms of causality. If his position is to be free of such circularity, the advocate of DN must take laws to be mere concomitances. In his view, the law that all metal expands when heated *is* the regularity that, in the eyes of a non-Humean, merely expresses that law. In general, for the advocate of DN, laws are regularities—nothing more.

But in that case, *any* true universal generalization expresses a law. YP ("everything is such that it's either yellow or it isn't a pencil on JMK's desk") describes a regularity. It correctly states that nothing has the property of being a pencil on JMK's desk on March 11, 2009, unless it also has the property of being yellow. But it doesn't express a law.

If laws are identified with regularities, as opposed to the structures that (so non-Humeans think) underlie them, then DN ends up having the same problems as Hume's analysis of causation. People often know what caused what, even though they couldn't for the life of them identify the relevant generalizations. It often happens that, *in order* to identify the relevant generalization, one must know what caused and what—that one must *already* have an explanation of the event in question, if only a crude one, if one is to identify the relevant generalization. Jones wants to know why the car blew up. He thinks about the sequence of events that he just witnessed. Smith got into his car and turned the key in the ignition; the engine turned over; a moment later the car exploded. If Smith falsely believes that Brown used his telekinetic powers to blow up the car, Smith's search for covering laws will be futile. It is only if Smith *does* have a correct explanation, if only an extremely low-resolution one, that he can even begin to search in a meaningful way for the relevant law.

Laws are hyper-explanations. (In this context, "law" and "statement that

expresses a law" are used interchangeably.) They are explanations of the explanations. First there are the *anomic* explanations. ("Nomic" means "involving or having to do with natural law." "Anomic" means "not involving natural law.") The explanations that we come up with in everyday life are anomic—"heating it caused it to expand," "falling caused him to break his leg," "drinking arsenic caused him to die," "being made of pinewood caused it to float when placed in water." The laws explain these explanations. "There is an underlying structure that guarantees that a thing will expand if it's made of metal that is being heated/that a person will break his leg as a result of such and such a fall/that a person dies if he consumes arsenic/that a thing floats if it's pinewood in water." Hume and Hempel believe that the event is to be explained in terms of the law; they believe that laws are logically prior to events. The events are to be understood in terms of the laws. But it's the other way around. To be sure, once the law is in place, new and better explanations are possible. But those higher-level explanations stand on the shoulders of lower-level ones that model the data without the help of statements describing laws. Hume and Hempel believe that it is on the basis of their *knowledge* of laws that people explain events. It's the other way around. People *already* have explanations—albeit low-level, low-resolution ones—of the phenomena; and *on that basis* they posit the existence of laws. Hypotheses concerning laws are distillations of existing, low-level explanations.

In conclusion, DN is viciously circular and is therefore a failure.

7.5 Explaining explanation

One principle always underlies the search for answers (here I am referring exclusively to attempts to answer *empirical* questions—questions concerning the spatiotemporal):

(MC[334]) causal anomalies must be minimized.

Everything will be explained when and only when it is known, for any given state of affairs, what caused that state of affairs to obtain. If you know all the causes, you know all the explanations.

"But surely that isn't true," it will be said. "I can know *what* killed Tim (he ate fish), but not know *how* it did so. (Was the fish rotten? Was it poisoned? Was Tim allergic to fish?)" Yes, but knowledge of the how is knowledge of the what. How-knowledge is high-resolution what-knowledge. If all you know about the cause of Tim's death was that it was brought on by his eating fish, then you don't *really* know what caused his death. Your not knowing the "how" consists in your not knowing the specific mechanisms involved; and your not knowing those mechanisms consists in your not knowing the "what."

That having been settled, let us turn to the question: how are causal anomalies minimized? The answer:

(ED[335]) the elimination of causal anomalies consists in the elimination of discontinuities.

You see a red ball roll in one end of a tunnel with opaque walls A moment later, you see a similar ball roll out the other end. Given this information, it is natural to suppose that

(H_1) the ball you saw entering the tunnel rolled from one end of it to the other and then exited it.

It's a theoretical possibility that H1 isn't the right explanation. Maybe what you saw is part of a magic-trick. But other things being equal, H_1 is almost certainly correct. And even if it isn't correct, it's obviously a very reasonable hypothesis. The reason is that it eliminates the discontinuities in the events that you observed by supplementing them with events that you did not observe. Your observations tell you that some ball was instantaneously annihilated and that some other ball instantaneously popped into existence a few feet away. But if H_1 is correct, none of these *apparent* discontinuities is an actual discontinuity. At least part of what makes H_1 a good hypothesis is that, to the extent that it's correct, reality is continuous. This suggests that to explain is to eliminate discontinuities. Explanation is discontinuity-elimination.

In response to this, it will be said that:

(OH)[336] We can use your own example to prove that, contrary to what you are alleging, explanation is not the same thing as discontinuity-elimination. A ball rolls into one end of a tunnel and a similar ball rolls out the other. But the second ball isn't identical with the first. The first was obliterated the moment it was completely inside the tunnel. At that instant, an iguana suddenly materialized in the place previously occupied by the ball. That iguana ran from one end of the tunnel. But just before it would have emerged from the other end of the tunnel, the iguana hit a force field. Because of that collision, the iguana turned into a ball just like the one that entered the tunnel and, in addition, that ball rolled out of the tunnel.

Let H_2 be this hypothesis. H_2 eliminates each of the discontinuities eliminated by H_1. But whereas H_1 is a good hypothesis, H_2 is rubbish. So you're wrong: to explain is *not* to eliminate discontinuities.

OH is misguided. Supposing that H_2 is correct, a red ball suddenly vanishes and is replaced with an iguana that, although continuous with the ball in respect of its *position,* is discontinuous with it in respect of its morphological, chromatic, and chemical properties; and, a moment later, that iguana suddenly vanishes and is replaced with a ball that, although continuous with the iguana in respect of its position, is discontinuous with it in respect of its shape, color, and chemical properties. Thus, merely H_2 replaces observed discontinuities unobserved discontinuities. H_1 doesn't posit any of the discontinuities posited by H_2. So contrary to what OH says, there are many discontinuities that H_1 eliminates that H_2 fails to eliminate. Moreover, there discontinuities that H_2 creates that H_1 does not create. These facts are not only consistent with, but strongly support, our thesis that explanation is discontinuity-elimination.

That said, there is a gap in our defense of that thesis:

(OH*) Clearly erroneous hypotheses can do just as good a job as correct ones of eliminating discontinuities. Here is an example. You see a ball roll into an opaque tunnel. Immediately upon entering the tunnel, the ball quickly but *continuously* turns into an iguana. The iguana runs to the

other end

of the tunnel. Once at the other end, the iguana quickly but *continuously* turns back into a ball; and that ball rolls out the other end of the tunnel.

Let H_3 be this hypothesis. Even though it's obviously unacceptable, H_3 does as good a job of eliminating discontinuities as H_1. Thus, contrary to what you are saying, discontinuity-elimination cannot be the essence of explanation.

Contrary to what OH* alleges, H_3 does *not* do as good a job as H_1 of eliminating discontinuities. First of all, even though it seems, at least initially, to do as good a job as H_1 of eliminating spatiotemporal discontinuities, H_3 creates a number of *theoretical* discontinuities that H_1 does not create. There is no independent evidence that balls turn into iguanas or *vice versa.* So, in supposing that H_3 is correct, we're supposing *either* (i) that the physical laws governing this one situation are very different from those governing other situations *or* (ii) that the observational data that we've had thus far, which strongly suggests that iguanas *don't* turn into balls or *vice versa,* is very misleading as to the nature of reality. If (i) is the case, the laws of nature, and therefore the very structure of reality, had to undergo an abrupt and unprecedented change in order for the changes described by H_3 to be possible. (And given the reasonable, though debatable (not to mention vague), supposition that the laws of nature are veritable spatiotemporal structures, it follows that H_3 fails to eliminate discontinuities in a particularly egregious way.) If (ii) is the case, there is so little conformity between the data of observation and the structure of reality that we have no good reason to expect any of our observation-based beliefs, including our belief in (ii) itself, to be correct. Given obvious extensions of these points, it's clear that H_2 creates similar theoretical discontinuities.

But H_1 doesn't have any of these problems. It doesn't create any of the theoretical (or spatiotemporal) discontinuities that either H_2 or H_3 creates, and it eliminates each of the discontinuities eliminated by either. All of which suggests that, just as we proposed, discontinuity-elimination is the essence of explanation.

8.0 Probability an epistemic concept

At time t, a prominent medical researcher says:

(S_1) "it's probable, but not certain, that Socrates had four kidneys."

> Given all of the data available up until time t, S_1 is the rational thing to say.
>
> More data becomes available. The researcher considers it. And, at time t*, that same researcher says:

(S_2) "Socrates only had two kidneys."

> Given all of the data available up until time t*, S_2 is the rational thing to say.

Question: Given that S_2 is true, does it follow that S_1 was false? No. Given the data available to the researcher at t, it *was* probable that Socrates had four kidneys. Like all probability statements, S_1 describes a relationship (one of bearing) that holds between some body of data and some hypothesis. S_1 does *not* describe some state of affairs. If it did, it would be to the effect that Socrates had two kidneys *to a certain degree,* in which case it would be meaningless.

8.1 Probability a property of statements

Probability is a property of statements, not of states of affairs.[337] It's a property of statements to the effect that a certain hypothesis is given a certain amount of support by a certain body of data. So it's a property of statements that describe bearing-relations. True probability statements correctly describe such relations; false ones incorrectly describe them.

An illustration: Smith tells his friend Brown that he's going on a tour of Europe. But he's vague about where he's going to go, and when. But he's not *completely* vague; he gives Brown some relevant information. A month after Smith leaves, Brown tells a friend of his that:

(PS_1) It's probable that Smith is in France right now.

PS_1 is correct. It's correct because, given the data available to Brown, it's far more reasonable to believe that Brown is in France than it is to believe that he isn't. And PS_1 is correct even if, at the time, Brown is Spain. It doesn't matter whether Brown is in France or not. What matters is that, given the available data, it's more reasonable to believe that he is than it is to believe that he isn't.

PS_1 says that, *given* such and such data, it's probable that Smith is in France. It's irrelevant that, from a phonetic viewpoint, there's no expression of the form "given such and such data." From a logical standpoint, there is such an expression (unless PS_1 is completely meaningless).

In general, probability is inherently relational. For this reason, the words "P is probably the case" may be correct when they come out of one person's mouth and false when they come out of some other person's mouth. This will happen if the data available to the one person makes P more belief-worthy than not-P and the data available to the other makes not-P more belief-worthy than P.

A complete description of the universe would contain no probability statements. Given any statement S (that isn't a probability statement), such a description would contain either S or not-S. Thus, probability is but "a measure of ignorance," as Laplace put it. Somebody who knew everything could never correctly describe a proposition as probable but not certain.

8.2 Probability not subjective

Probability is a mind-independent logical relationship holding among statements. The degree to which a given body of data D confirms a given statement S has nothing to do with anyone's opinion about anything: D confirms S to that degree whether or not anyone knows it—whether or not anyone exists. Probability statements are *useful* only to the ignorant. (An omniscient being wouldn't have to bother with them.) But that doesn't mean it's only insofar as people are ignorant that they're true. It means the opposite. They're useful to the ignorant because, when correct, they describe relations that actually hold between what (little) they know and what they'd like to know.

Given a certain body of data, it's more reasonable to believe P than it is to believe not-P, even though, given that same data, it is *not* reasonable to believe P. To believe P is to be certain that P is correct. The data one has may make P *more* worthy of belief than not-P without making P worthy of belief. Given the data available to me, it's more reasonable to believe that my friend Sally is in Washington than it is to believe that she isn't. But since, given that data, it's only marginally more likely than not that she's in Washington, it would be irrational of me to *believe* that she he is.

8.3 Is there such a thing as non-epistemic probability?

According to some, probability is sometimes a property of events, as opposed to statements. Here's the idea. At time t, electron e jumps to a higher orbit. Though the laws of nature didn't require it to jump, or not to jump, at that time, they did *dispose* it to jump. Here, it is said, we are dealing with a case of intrinsic, as opposed to evidential, probability. Supposing that one knew everything there was to know *other* than the fact that e jumped orbit at t, one could not on that basis predict with any certainty that e would jump at that time. Under these circumstances, it might be said, it *would* be correct to say that, for reasons having nothing to do with ignorance on anyone's part, the statement "it is more probable than not that e will jump at t" is correct.

For argument's sake, we'll suppose that there are indeed undetermined events—events that don't have to happen but do.

One could argue on the following grounds that intrinsic probability is a kind of evidential probability. Let S be a true statement alleging the occurrence of event E. S has an "intrinsic" probability of greater than 50% iff S_1. . . . S_n make it more likely than not that S, where S_1 . . . S_n is a list of every true statement *that doesn't* logically entail S. The idea is that, if you knew everything there was to know *other than the fact that E jumps orbit at t,* you could not learn more that was relevant to whether E would jump without actually learning whether E would jump. But, it will be said, the story is very different with statements that, as a matter of natural law either must happen or cannot happen. Let S* be a true statement falling into this category. No matter how small the likelihood that S* is false, there are true statements that,

if conjoined with the truths already known, will shrink that likelihood even further *without eliminating it.* (It may be that, because of interference effects, it's impossible to acquire more information that is relevant to whether S* is correct or not without actually learning whether S* is correct or not. But that would only be because those interference effects made it impossible to learn the truths that would enable you to shrink your margin of doubt.) Thus, intrinsic probability is a kind of evidential probability. One is dealing with intrinsic probability when, setting aside problems relating to interference effects, there *is* a limit to how much, short of learning whether the target statement is true or not, one's margin of error can be reduced, and one has non-intrinsic evidential probability if there is no such limit.

Nonetheless, I'm not convinced that intrinsic probability is, in fact, a form of evidential probability. As before, let S be a true statement alleging the occurrence of some event E that has an intrinsic probability of greater than 50% but less than 100%, and let S_1 . . . S_n be a list that contains every true statement other than those that entail S. The statement "given S_1 . . . S_n, S is more probable than not" does not say that, given S_1 . . . S_n, we're stuck with more questions than answers if we deny S. Were it supposed true that *Hamlet* was written by a levitating cat that, despite being incapable of self-consciousness, was deeply neurotic, all manner of beliefs concerning physics, biology, and psychology would have to be revised. But, assuming the truth of S_1. . . S_n, it doesn't much matter whether S is accepted or rejected. It is, by hypothesis, a free wheel. I must therefore conclude that intrinsic probability, if it exists, is not evidential probability.

I leave it to the reader as an exercise to say why intrinsic probability is not, contrary to what some might hold, identical with statistical probability; and I also leave it to the reader to say what exactly intrinsic probability is.

8.4 What is it for one statement to confirm another?

In this section, we'll put forth some more reasons to think that explanatory (evidential) probability is not statistical probability. It will help if we start with a summary of our findings thus far.

Other things being equal, a hypothesis is to be preferred to its negation if

and only if more causal anomalousness is created by rejecting it than by accepting it. The degree to which a hypothesis is to be preferred to its negation is directly proportional to the degree to which it does a better job than the other of minimizing the net amount of causal anomalousness in the world.

Probability is not a statistical notion. The word "probability" has a statistical meaning. This is the meaning that it has in the statement: "the chances of a person who's picked at random having red hair are one in eight." All this means is that one person in eight has red hair.

But this isn't the sort of probability that is relevant to explanations. When you say ("SNS" is short for "Smith no show"):

(SNS) "given that Smith hasn't shown up yet, he's probably not going to show up at all,"

you are *not* making a statistical statement; you are not saying that in more than 50% of situations like the present one Smith doesn't show up. SNS could be true even if Smith has never failed to show up to anything in his life. And it could be false even if, in the past, he's failed to show up to 90% of his appointments.

Statements affirming the probability of hypotheses seldom even *appear* capable of statistical interpretations. You can know that x is likely to do y in circumstances z without ever having seen x, or anything much like it, do y, or anything much like, in circumstances z, or anything much like them. For example, we have a pretty good idea of what at least some of the consequences of detonating a hydrogen bomb over the South Pole would be, even though nothing much like that has ever occurred.

"But that's where you're wrong," contemporary Humeans say. "You see, it's only if it's interpreted in an absurdly narrow way that that the statistical conception of explanation has any of these problems." And they defend this position by saying that ("SE" stands for "statistical explanation"):

> (SE) If a hydrogen bomb were detonated over the South Pole, it would by no means be unprecedented. True, we haven't detonated hydrogen bombs *over the South Pole.* In fact, we haven't detonated *hydrogen* bombs anywhere. But we've detonated weapons *similar* to hydrogen

bombs over places *similar* to the South Pole. Given this information, we have a pretty good idea of what the consequences would be of detonating a hydrogen bomb over the South Pole.

And while you're right that SNS may be correct even if Smith has never missed an appointment in his life, you're wrong to conclude that explanation is not to be understood in statistical terms. If one is to be justified in accepting SNS, one doesn't need to know that *Smith* ever failed to show up to anything; and one doesn't need to know that what the relevant individuals, whoever they might be, failed to do was to show up. One need only know that people similar (in some respect relevant in the present context) to Smith failed to do something similar (same qualification) to showing up. One need only know that people *like Smith* (e.g., people who have serious drug problems) failed in some way that would be comparable to Smith's failing to show up (e.g., those people did keep their promises).

It's easy to ridicule the statistical conception of explanation by showing that events *exactly like the one being explained* haven't occurred. But the situation being explained needn't be a perfect replication of any past situation. It need only be *like* past situations in some very abstract respect.

The problem with this response is that there is no way to understand the concept of "relevant similarity" in statistical terms; and in deciding what is relevantly similar to what, the advocate of SE is drawing on hypotheses that can't possibly have received any confirmation of a purely statistical kind.

Any two things are similar in some respects and dissimilar in others. Some of these similarities are explanatorily significant, and others are not. It often happens that, despite our not having first conducted context-specific research, we can distinguish the relevant from the irrelevant forms of similarity. For example, it's obvious that, when trying to figure out what the consequences would be of detonating a given bomb over the South Pole, it doesn't matter whether that bomb is green as opposed to red, but it *does* matter whether it's a nuclear as opposed to a non-nuclear bomb. So it's obvious that a red hydrogen bomb would be much more similar, in the respects relevant in this context, to a green hydrogen bomb than it would be

to a red non-nuclear bomb.

But this wouldn't be obvious at all if our evidence for it were strictly statistical. Given that, for any two things, one can contrive some respect in which they are similar, it follows that, with a few trivial qualifications, anything can confirm anything to any degree.

This is what Nelson Goodman's argument shows. (See Section 7.3). Find *a* respect in which green emeralds are similar to blue emeralds, and the fact that, thus far, all emeralds have been green confirms that, after this point, they'll all be blue. (All green emeralds examined before the present are similar to blue emeralds examined in the future in that they're all "grue.")

Another example: Find *a* respect in which herbivorous brontosauruses resemble cars that aren't affected in any adverse way by high-speed collisions, and the fact that all brontosauruses were herbivorous confirms that a speed collision won't damage your car. (Herbivorous brontosauruses are similar to cases of your crashing your new, as of yet unscathed, car in that they are both "carbivorous." Something is "carbivorous" if it's either an herbivorous brontosaurus or a car driven by you after the present time that's involved in a high-speed collision and isn't damaged by it.")

We know that terms like "grue" and "carbivore" pick out explanatorily hollow properties, and that's why we don't base our driving habits on our beliefs about the dietary habits of dinosaurs. We know what to infer and what not to infer *only* insofar as we know what kinds of similarity count. A *precondition* for making the right inferences is knowing what the relevant kinds of similarity are. Thus, there's no way that judgments as to what the relevant forms of similarity are could have a strictly statistical basis. (That said, they obviously can and do have a *partly* statistical basis. But that's not enough to make a go of the statistical model.) The concept of probability is to be understood in terms of the concept of explanatory integrity, not that of statistical frequency. The concept of explanatory integrity is to be understood in terms of the concept of *causal* integrity. The concept of causal integrity is to be understood in terms of the concept of continuity of change. One very important kind of change is change of position (motion). But not all change is change of position. Changes in temperature aren't changes in position. Nor are changes in color, shape, or structure. That said, at least some of these

forms of change are to be understood *in terms of* the concept of change of position. But the same can't be said of other explanatorily important kinds of change. It can't be said of higher-order changes (changes in the manner of change), since higher-order changes aren't *spatiotemporal* changes at all. (A second-order change is a change in a spatiotemporal change; it is not itself a spatiotemporal change, and the same is obviously true of all other higher-order changes.) Earlier we considered some clearly erroneous hypotheses that seemed at first to do just as good a job of preserving continuities as some clearly worthy hypotheses. Had these initial impressions been correct, our analysis of explanation would have been in trouble. But they weren't correct. When we took a second look, we found that the clearly wrong hypotheses did a much worse job than their meritorious counterparts of dealing with higher-order continuities. And the continuities in question were ones that we already knew to be explanatorily important; they weren't contrived ad hoc to buttress our theory.

9.0 The raven paradox

Because they don't grant the existence of causality, except of a degraded and useless kind, Humeans are categorically incapable of producing viable theories of confirmation. One sign of this is that Humean theories of confirmation always have paradoxical consequences.

The most famous of these famous paradoxes is the so-called Raven paradox. This was discovered by C.G. Hempel[338], who was himself a hardcore Humean. We've already considered Hempel's unsuccessful, Humean analysis of explanation, and we'll soon consider Hempel's equally unsuccessful, Humean attempt to solve the raven paradox.

In any case, here is that paradox:

(RP) Logically equivalent sentences are confirmationally equivalent. This means that, if P and Q are logically equivalent, then not only does evidence E confirm P just in case it confirms Q, but E confirms P *to degree n* just in case it confirms Q to that same degree.

The sentence:

(1) "All ravens are black" is logically equivalent with:

(2) “All non-black things are non-ravens.”

The existence of non-ravens (e.g., pianos, pants, carpet fibers) that are pink or green confirms (2), but does *not,* unless our intuitions are very wrong, confirm (1). So, impossible though it must be, (1) and (2) are logically, but not confirmationally, equivalent.

An analogous paradox is associated with each of many others pairs of logically equivalent sentences. For example:

(i) “if it’s made of metal and is being heated, then it’s expanding.”

is logically equivalent with:

(ii) “if it isn’t expanding, then it isn’t both made of metal and being heated.”

(ii) is confirmed by the fact that the non-expanding Tupperware in my freezer is not made of metal (or a fortiori of metal that is being heated). But, unless our intuitions are very wrong, that fact provides no support for (i).

Some, including Hempel himself, have argued that, our intuitions notwithstanding, the existence of pink pianos *does* confirm (1) to degree n exactly if it confirms (2) to that same degree and that the existence of non-expanding Tupperware in freezers does confirm (i) exactly as much as it confirms (ii). Others have argued that the paradox should be solved by advocating non-standard logics, relative to which logical equivalence doesn’t guarantee confirmational equivalence.

We’ll evaluate these solutions after we put forth our own. It’s clear that, whatever merits these solutions have, they are counterintuitive. And the very idea of logical equivalence not guaranteeing confirmational equivalence is of questionable coherence. But even if it can be shown that an internally consistent logic could be invented relative to which logical and confirmational equivalence pull apart, it seems excessive to throw out classical logic to solve this one problem. Fortunately, the paradox can, I will argue, be solved without going to such lengths.

9.1 Inductions by enumeration identical with

crypto-inferences to the best explanation

In this context, (i) may be treated as saying the same thing as:

(MH) For any x, if x is made of metal and x is being heated, then x is expanding.

And (ii) may be treated as saying the same thing as:

(HM) For any x, if x isn't expanding, then x isn't made of metal that is being heated.

MH can be taken *either* to express an "accidental generalization" *or* to express a principled relationship. Taken in the first way, it does nothing more than affirm a purely de facto concomitance of two properties (viz. the property of being heated metal and the property of being something that expands). Taken in the second way, MH seems to say that all heated metal expands, *the reason being* that a thing's being heated metal provides some kind of nomic basis for its expanding. In this section, we will focus on the first, strictly extensional, reading. In the next, we will focus on the second, nomically committed reading.

> Let us paraphrase MH, when it is taken as a purely de facto generalization, as follows:

(MHA)[339] For any x, if x is made of metal and is being heated, then, by happenstance, x is expanding.

So, in this context, take the operator "by happenstance" to be short for "not because x's having the first property provides a causal or nomic basis for its having the second." (By "the first property," I mean the property mentioned before the "then," viz. the property of being made of metal that is being heated; and by "the second property," I mean the property referred to after the "then," viz. the property of expanding.) We don't need to worry about whether this corresponds to the usual meaning of "as a matter of happenstance," since we are using this expression in a stipulative manner.

If we take MH in this way—if we take it to mean the same thing as MHA—does Hempel's paradox arise? No. Let phi and psi be any two properties such that *x has psi* isn't an analytic consequence of *x has phi,* or *vice versa.* If I know that there is no causal or, more generally, nomic connection between

x's being a phi, on the one hand, and x's being a psi, on the other, then I have no reason to infer, from x's being a phi, that x is also a psi. Even if, thus far, every ball withdrawn from the urn has been black, unless I have reason to believe that this concomitance has some basis in natural law or in some causal mechanism (e.g., the presence of a force field surrounding the urn that allows entry only to black balls, or some person's resolute intention to allow no non-black balls into the urn), I must ipso facto regard it as nothing but coincidence that, thus far, every ball withdrawn has been black; and I must therefore *not* think it more likely than not that the next ball withdrawn will be black—for to think otherwise would be to commit the Monte Carlo fallacy. [340] Thus, supposing that, in fact, this concomitance is mere coincidence, the generalization *all the balls in the urn are black* fails to receive any positive confirmation from its instances.[341] For the same reason *mutatis mutandis,* the statement *all non-black things are not balls in the urn,* unless read as having nomic force, fails to receive support from *its* instances.

And, of course, it also follows that HM, if read as doing nothing more than affirming a de facto concomitance, doesn't receive support from its instances. Thus

(HMA) For any x, if x isn't expanding, then, by happenstance, x isn't made of metal that is being heated,

doesn't receive support from the fact that the non-expanding Tupperware in my freezer isn't made of metal that is being heated.

So, if MH and HM are read as having no nomic content—if they are read as doing nothing but expressing de facto concomitances—then, once it is seen that de facto generalizations don't receive support from their instances, HM doesn't even *appear* to receive support from the fact the non-expanding Tupperware in my freezer isn't made of heated metal, and MH doesn't even *appear* to receive support from the fact that the metal pan I put in the oven is expanding. A fortiori the former fact doesn't even *appear* to confirm HM more than it does MH, and the latter fact doesn't even appear to confirm MH more than it does HM.

And we therefore don't get the kind of (apparent) confirmational non-equivalence that gives rise to Hempel's paradox. For that paradox depends

precisely on its being the case that, despite the (alleged) fact that HM and MH are logically equivalent, they are not, or at least appear not to be, confirmationally equivalent. And they appear to be confirmationally non-equivalent only in so far as (e.g.) the existence of a heated, expanding metal objects confirms MH but not HM, and only in so far as (e.g.) the existence of non-expanding, non-heated, non-metal objects confirms HM but not MH. But once it is made clear that de facto generalizations don't receive support from their instances, it follows that, if MH and HM are read as free of nomic content, nothing even appears to give more support to the one than to the other. And the paradox fails to arise.

9.2 Inductions by enumeration identical with crypto-inferences to the best explanation (continued)

So, if Hempel's paradox is to arise, at least one of MH and NH must be taken as doing more than affirming a de facto concomitance. And it seems that, thus taken, MH is to the effect that all heated metal expands *for the reason that* a thing's being metal that is being heated provides some kind of nomic or causal basis for its expanding. Let us paraphrase MH, when it is taken in this way, as follows:

(MHN)[342] For any x, if x is made of metal and is being heated, then, as a matter of principle, x is expanding.

So, in this context, take the operator "as a matter of principle" to be short for: "because x's having the first property provides a causal or nomic basis for its having the second." Echoing what we said earlier, it is irrelevant this may not correspond very well to the usual meaning of "as a matter of principle," since we are here using this expression in a stipulative manner.

At this point, we have a choice. We can take HM as having no nomic content—as taking it to have the same import HMA. Or we can take it to have nomic content. Let's consider both options, starting with the first.

Given that MH is taken to do more than affirm a de facto concomitance, if NH is taken to affirm nothing but a de facto concomitance, the two will, quite trivially, not be equivalent. For obviously MHN is not equivalent with

HMA. The first is committed to its being the case that *inter alia* a thing's being made of heated metal provides some kind of nomic basis for—or, in any case, is nomically or causally connected in some way with—that thing's expanding. The second isn't committed to there being any kind of nomic, or otherwise not purely extensional, connection between the property of non-expanding and the property of being made of heated metal. MHN is true only if certain nomic connections hold; HMA can be true in the absence of those connections. So there is no equivalence, or even the appearance of equivalence for that matter.

We've just seen that, if we take MH to have nomic force (which we must —see above—if Hempel's paradox is to arise), Hempel's paradox doesn't arise if we *don't* take HM to have nomic content. In other words, if we take HM to do nothing more than affirm a de facto concomitance, then it won't, as we've just seen, even appear to be equivalent with MHN, and nothing even resembling Hempel's paradox arises. So if that paradox is to arise, then we must take HM to have nomic content.

In light of this, consider the sentence:

(HMN) For any x, if x isn't expanding, then, as a matter of principle, x isn't made of metal that is being heated,

HMN says that a thing's not being an expander provides some kind of (partial) nomic basis for its not being heated metal. Of course, the relationship that HMN bears to HM is analogous to the relationship that MHN bears to MH. But what is most important here is that HMN would be the right way—or, in any case, the (or, at least, *a*) natural way—to take HM *if* the latter were taken to do more than affirm a de facto concomitance between two properties.

To clarify this last point, suppose that you are speaking to a friend who is about to drink a large quantity of arsenic but who doesn't realize that consuming arsenic in large quantities leads to death. You tell him, warningly, "people who consume large quantities of arsenic die soon thereafter." Here you are obviously not affirming a purely extensional connection between, on the one hand, the property of being a person who drinks large quantities of arsenic at a given time and, on the other hand, the property of being something that dies soon thereafter. What you are saying is that a person x's

drinking large quantities of arsenic at a given time *leads to,* or in any case provides some kind of nomic foundation for, x's dying soon thereafter, *that* being why there is a concomitance between the property of being a person who drinks arsenic, on the one hand, and the property of being dead soon thereafter, on the other. That this is the right interpretation is made clear by the fact that you would be *failing* to warn your friend not to drink the arsenic if you had instead said: "all people who have drunk arsenic died soon thereafter. But that had nothing to do with their drinking arsenic. You see, it turns out that the only people who've ever drunk arsenic belong to religious sect X; moreover, all of those people were about to be executed. One of the precepts of X is that, if you are about to be executed, you must drink arsenic in order to get into heaven. And that's why, thus far, each person who's consumed arsenic died soon thereafter.

"Just as the most natural way to take "people who drink arsenic die"/"heated metal expands" to do more than affirm a de facto concomitance is to take it to mean that a person's drinking arsenic/being heated metal provides some kind of causal or nomic basis for that person's dying/expanding, so the most natural way to take HM to have nomic force is to take it to mean that a thing's being a non-expander provides some kind of causal or nomic basis for its not being heated metal. And that is precisely the reading of HM that is embodied in HMN.

HMN is clearly *not* equivalent with MHN. Suppose, if only for argument's sake, that a thing's being made of heated metal *causes* it to expand. So suppose that:

(MHC) "as a matter of natural law, x's being heated metal *causes* x to expand."

That would certainly be sufficient for the truth of MHN. In light of this, consider the statement that:

(HMC) "as a matter of natural law, x's not expanding causes x to not be metal that is being heated."

Of course, MHC, *if true,* would be sufficient for the truth of HMN. But MHC is *not* equivalent with HMC. Suppose it is a law that **x's being a human who drinks arsenic at t causes x to be dead at some later time t*.**

This very reasonable and (if refurbished a bit) approximately correct statement obviously doesn't entail, and therefore isn't equivalent with, the false and practically incoherent statement that *x's not being dead t* causes x to be a human who doesn't drink arsenic at earlier time t.*

For similar reasons, MHC is not equivalent with HMC. So, since MHC and HMC are sufficient for the truth of, respectively, MHN and HMN, it follows that MHN and HMN are not equivalent: they don't hold in the same models, the same possible worlds. In fact, whereas MHN, or something similar to it, *does* hold in our world, the same doesn't go for HMN or anything at all like it.

9.3 Inductions by enumeration identical with crypto-inferences to the best explanation (continued)

We've seen, first, that MH must be read as having some kind of nomic content if Hempel's paradox is to arise (i.e., if there is to even be the appearance that MH is logically, but not confirmationally, equivalent with HM). And we've also asserted that, if MH read as having nomic force, its meaning is given by MHN, and we made the same assumption *mutatis mutandis* about HM.

But, it must be admitted, we have provided no argumentative basis for this assertion, our grounds for it being purely intuitive. Which brings us to our next point, namely, that there is another possible nomic reading of MH:

(MHW[343]) "thanks to the existence of some kind of nomic (or causal) connection between the two properties about to be mentioned [viz. the property of being heated metal, on the one hand, and the property of expanding, on the other], for all x, if x is heated metal, then x is expanding."

How is MHN different from MHW? MHN says pretty much the same thing as:

(*) for all x, if x is heated metal, then x is expanding, and this is because x's being heated metal provides a nomic basis for x's expanding.

By contrast, MHW says that:

(^) for all x, if x is heated metal, then x is expanding, and this is because *either* x's being heated metal provides a nomic basis for x's expanding *or* x's expanding provides a nomic basis for x's being heated metal.

To be sure, the expression "provides a nomic basis for" is an obscure one, which I've done little or nothing to clarify; and it might therefore seem that there is no difference between (*) and (^). But, obscure though that expression may be, (*) and (^) are demonstrably different, and non-equivalent, statements.

It's true (or so let us suppose, for argument's sake) that all people who consume large quantities of arsenic die soon thereafter. And this is not an accident: human biology being what it is, the chemical properties of arsenic being what they are, there's a nomic connection between consuming arsenic and dying. Vague as the expression "provides a nomic basis for" is, would it be correct to say that being a dead human at 3:10 P.M. "provides a nomic basis for" being a person who drinks arsenic at 3:00 P.M.? More generally, supposing that x's having psi at t* is an *effect* of x's having phi at earlier time t, does it make sense to say that x's having psi at t* provides a nomic basis for x's having phi at t? No. My being obese provides a nomic basis for my having a heart-attack. But surely my having a heart attack doesn't provide a nomic basis for my being obese. And that's all we need to know in this context about what it means to say that one thing provides a "nomic basis" for another. So let us proceed.

Suppose that, in possible world W, all things made of heated metal expand. Further suppose that, in W, x's being something that is expanding causes x to turn into something that is metal whose temperature is continuously rising and whose volume is constantly increasing. And, to make things simple, suppose that, in W, all things made of heated metal got that way by, at some point in their histories, being things that were expanding. In that case, MHW would be correct in W: for there would, in that world, be a nomic connection between a thing's being heated metal, on the one hand, and its being something that expands, on the other; moreover, the fact that, in W, all heated metal expands would have a basis in this nomic connection. But it would not, at least not on that account, be correct to say that, in W, a thing's being made of heated metal provides a nomic basis for its expanding. In W,

things expand *before* they become expanding heated metal. So, in W, a thing's being heated metal obviously cannot *cause* it to expand.

But there is a further point: in W, a thing's expanding is not to be *understood in terms of its being heated metal.* It is the other way around. In W, a thing's expanding is causally *and explanatorily* more basic than its being heated metal. In any case, just as (see above) it would be wrong, or at least very odd, to say that being dead at t* provided a *nomic basis* for drinking arsenic at earlier time t, it would, for parallel reasons, be wrong to say that, in W, being heated metal provided a nomic basis for being a thing that expands. And this entails that, however vague the expression "provides a nomic basis for" is, * and ^ are, very clearly, *not* equivalent.

9.4 This argument concluded

First of all, MHW obviously isn't logically equivalent with HMA: the former does, whereas the latter does not, allege the existence of certain nomic dependencies. More importantly, MHW isn't equivalent with:

(HMW) "thanks to the existence of some kind of nomic (or causal) connection between the two properties about to be mentioned [viz. the property of not expanding, on the one hand, and the property of not being heated metal, on the other], for all x, if x is not expanding, then x is not heated metal."

Given only that there is a nomic connection between a thing's having phi and its having psi, it doesn't follow that there is a nomic connection between a thing's *not* having psi and its *not* having phi. There is a causal, and therefore a nomic, connection between Smith's falling off of tall building X at t, on the one hand, and his being dead at later time t*, on the other. Smith's falling off of X is an event. It is something that makes things happen (e.g., various bones of Smith's being shattered); and, what is probably a more general notion, it is something in terms of which things can be explained.

Compare Smith's situation with Brown's. Brown does *not* fall off X, or any other building, at t; and Brown is alive and well at t*. Obviously Brown's *not* falling off X at t doesn't *cause* Brown to be alive at t*. More generally, Brown's being alive at t* is not, except perhaps in some extremely derivative sense, to be explained in terms of Brown's not falling off of X. Whereas

Smith's falling off of X is an event, Brown's *not* falling off of X is *not* an event, and it doesn't consist of events. A chair consists of events (various miniscule displacements of mass-energy), and does a corporation (various actions on the part of sentient beings); so, indeed, does any enduring entity—any structure, any "object," any non-instantaneous event. But Brown's *not* falling off X isn't identical with an event and isn't constituted by events.

Of course, since it's a truth, and therefore a "fact," that Brown did not fall off of X, we might say that the (negative) "fact" of his *not* falling off of X exists no less than the (positive) fact of Smith's falling off of X. But that doesn't undermine our point that Brown's *not* falling off of X isn't identical with, or constituted by, events.

While it is a matter of intense debate what nomic relations are, it is pretty clear they hold only among occupants of the space-time manifold. They clearly hold among events (my pounding the table causes it to break). Maybe they hold among things other than events. Maybe, for example, they hold among conditions. And maybe they hold among objects. But while it is unclear whether ***al*** spatiotemporal entities stand in nomic relations, it *is* clear that nothing *non*-spatiotemporal stands in them. My pounding the table is nomically connected to the table's breaking. But my *not* destroying Jupiter isn't nomically related to anything. My pounding the table is an event; it makes things happen; things can be explained in terms of it; it has a place in the tapestry that constitutes the spatiotemporal world. My *not* destroying Jupiter is nowhere in that tapestry, and therefore isn't fit to stand in nomic relations to anything. There is no event of my not destroying Jupiter. There is thus no condition or object that is constituted to any degree by that non-event. Therefore, it isn't identical with, or constitutive of, anything that falls within the scope of natural law. It isn't, therefore, nomically connected to anything.

A thing's being composed of metal that is of a certain temperature is, in the relevant respects, like my pounding the table. Things are to be explained in terms of it. And, though not itself an "event," it is *constituted* by various interacting events, all of which occupy space-time. So the skillet's having a certain temperature, at a given time, is something that can, and indeed obviously must, stand in nomic relations. And the same is true of the skillet's expanding during a given interval.

But a thing's *not* expanding is, in the relevant respects, like my *not* destroying Jupiter, and so is a thing's *not* being heated metal. So, whereas a thing's being heated metal is, or at least could be, nomically related to its expanding, it doesn't follow, and it isn't the case, that a thing's *not* expanding could be nomically related to its *not* being heated metal.

These points make it clear why MHW isn't logically equivalent with HMW. A thing's being heated metal *can* (and, in fact, must) be nomically related to other things, and the same is true of a thing's expanding. But a thing's *not* expanding, like my *not* destroying Jupiter, isn't nomically related to anything, and neither is a thing's *not* being heated metal. And that's why, even though, when read as merely affirming *de facto concomitances,* MH and HM are equivalent, they cannot always be intersubstituted *salva veritate.* In particular, they cannot be so intersubstituted in contexts governed by the operator that begins each of MHW and HMW (viz. "'thanks to the existence of some kind of nomic (or causal) connection between the two properties about to be mentioned"). There is nothing mysterious about this. Many operators don't permit intersubstitutions of logically equivalent expressions. Logically equivalent sentences can be truth-preservingly intersubstituted in contexts governed by (e.g.) "it is true that," but not, at least not categorically, in contexts governed by "Timmy believes that." The operator "there is a nomic connection between the facts expressed by the antecedent and consequent of the following statement" doesn't tolerate intersubstitutions of logically equivalent sentences. For, even if:

(T1) "there is a nomic connection between the facts expressed by the antecedent and consequent of the following statement: if x is heated metal, then x is expanding"

is true, it doesn't follow that

(T2) "there is a nomic connection between the facts expressed by the antecedent and consequent of the following statement: if x isn't expanding, then x isn't heated metal"

is also true. In fact, T1 is, for the reasons just given, quite plainly false.

9.5 Hume-internal solutions to the raven paradox

Quine's (1977) solution is to say that that logical equivalence doesn't guarantee confirmational equivalence. This idea is totally incoherent. P and Q are equivalent if they are true in exactly the same worlds. Thus, supposing that P and Q are equivalent, it immediately follows that, given any proposition R, the likelihood of R's being true in a world where P is true is identical with the likelihood of R's being true in a world where Q is true. Quine's proposed solution is thus a non-starter.

Hempel proposes a different solution. It obviously *seems* as though the existence of a pink piano provides less support for:

(1) "All ravens are black"

than does the existence of a black raven. And it obviously *seems* as though the existence of pink pianos provides more support for:

(2) "All non-black things are non-ravens'

than it does for (1). But, says Hempel, these intuitions are misleading. Under ordinary circumstances, if we know that something is a piano, we almost certainly *already* know that it isn't a raven.

But imagine the following. You have bet your friend Smith a $1,000,000 that all ravens are black. There is some object x such that, although you and Smith don't yet know it, x is a piano. All you know about x is that it's pink. In that case, when you *discover* that x is a piano, and thus a non-raven, you (but not Smith) will breathe a sigh of relief; and your confidence in (1) will surge, and so, of course, will your confidence in (2).

Now imagine a different scenario. You have bet your friend Smith a $1,000,000 that all non-black things are non-ravens. There is some object y such that, although you and Smith don't yet know it, y is black. All you know about y is that it's a raven. When you discover that y is black, you will breathe a sigh of relief; and your confidence in (2) will surge, and so, of course, will your confidence in (1).

Thus, to Hempel, the confirmational disparities between (1) and (2) are illusory. When we imagine ourselves trying to confirm (2), we assume that we are looking at things that we *already* know to be non-ravens, but we don't

make the same assumption *mutatis mutandis* when we imagine ourselves to be trying to confirm.

Hempel's solution is better than Quine's, but it's still no good. The problem is that, if it's known that there is no nomic or causal relationship between a thing's being a raven, on the one hand, and its being black, on the other, then there is ipso facto no reason to believe that the next raven you see will be black. Maybe you've seen a million ravens, all of them black. It's doesn't matter. Unless you have reason to believe that a thing's being a raven is connected somehow to its being black, you're guilty of the gambler's fallacy if, on the basis of that data, you believe that the next raven you see is (likely to be) black. But if you *do* believe that a thing's being a raven is connected in some way to its being black, then your confidence in (1) is derivative of your confidence in:

(1*) a thing's being a raven provides a basis of some kind for its being black.

(1*) isn't equivalent with (2*). We've already seen why. And we've already see why (1*) isn't equivalent with:

(2*) a thing's being non-black provides some kind of a basis for its being a non-raven.

If the existence of a black is raven is to provide any legitimate support for (1), there must be a causal or nomic connection between ravenhood and blackness. Once this is realized, two things immediately become obvious. First, Hempel's solution is spurious, since it presupposes that black ravens can confirm (1) *without* there being any connection of the sort just described. Second, (1) must be read as having nomic force. Since, thus read, it is no longer equivalent with (2), the paradox vanishes.

9.6 These points summed up

If read as doing nothing more than affirming de facto concomitances:

(MH) if x is made of heated metal, x expands

and

(HM) if x isn't expanding, x isn't made of heated metal

don't receive support—and, to anyone who isn't guilty of the gambler's

fallacy, don't even appear to receive support—from their instances and are therefore vacuously confirmationally identical. So, thus read, they are logically *and,* quite clearly, confirmationally identical; and Hempel's paradox doesn't arise. If they are read as doing *more* than affirming de facto concomitances, they must be read as alleging, or at least being committed to there being, some kind of nomic connection between the properties in question. But, no matter what nomic readings we give these sentences, they don't even appear logically equivalent. (This is, to put it very crudely, because logically equivalent sentences cannot, at least not in general, be intersubstituted *salva veritate* in contexts governed by nomic operators; e.g. "there is a nomic connection between the following statements.") Bottom line: once it is realized that accidental generalizations don't receive support from their instances, this being the prerogative of principled generalizations, it isn't possible for Hempel's paradox even to arise.

Chapter 18

Modality and Non-Existence

1.0 Necessity and possibility properties of statements, not objects

A modal statement is one that says what *could* be the case or what *must* be the case. Here are some examples of such statements:

(1) Triangles must have three sides.

(2) It's possible to make money without selling out.

(3) There's no way anyone could have survived that fall.

(4) Assuming that Bob has fifty cars, he must have at least some money.

(5) John's becoming stronger was an inevitable consequence of his lifting weights all summer.

For any proposition P, it's necessarily the case that P ***iff*** it's not possible that not-P; and it's possible that P iff it's not necessarily the case that not-P. ("Iff" is short for "if and only if.") It's necessarily the case that Jim is in the barn iff it's not possible that Jim is not in the barn; and it's possibly the case that Jim is in the barn iff it's not necessarily the case that Jim is not in the barn.

There are no impossible objects—no objects that couldn't possibly exist. And there are no possible but non-actual objects— and no objects that could exist but don't. Possibility and impossibility are properties of propositions, not of objects. For example,

(6) square circles cannot possibly exist

says that the proposition

(7) there are square circles

cannot possibly be correct. (6) doesn't make the self-contradictory claim that there exists some object x such that x is a square circle and such that x cannot exist. And

(8) happiness is a possibility for Jim, even though he's still miserable

says that, even though Jim is now miserable, the proposition

(9) Jim is happy

could be true. (8) doesn't make the absurd claim that there exists some x such that x is Jim's as of yet nonexistent happiness and such that x could exist.

1.1 Existence and non-existence properties of propositions, not of objects

The points just made concerning the concepts of possibility and impossibility are to be understood in terms of more general points concerning the concepts of existence and non-existence.

Nothing doesn't exist. There exists no x such that x doesn't exist. Statements that, given their surface-structures, appear to be attribute non-

existence to objects in fact attribute falsity to propositions. A story will make this clear.

An unscrupulous person claims to have been attacked. In reality, he wasn't attacked; but, for whatever reason, it's to his advantage for people to believe otherwise. He describes his non-existent assailant to a police sketch artist. The resulting sketch is distributed across the nation. If we say that the sketch depicts a nonexistent person, we're saying that there exists some person who doesn't exist who it depicts. But that would be absurd. The sketch is better thought of as saying: *there is somebody having such and such features,* these being the features that a person *would* have to have *if* he were to fit the sketch. Though false, that *statement*—that ***existence-claim*** obviously exists.

Similarly, the cartoons that brought Fred Flintstone and Bart Simpson into our homes can be thought of as making false statements of a similar kind. In the case of Fred Flintstone, that statement might be something along the lines of: there exists a boorish man who has pet dinosaur named "Dino," a wife named "Wilma," etc. In the case of Bart Simpson that claim might say something along the lines of: "there exists a bratty boy who torments his oafish father with his loveable shenanigans (etc.)" All of these *statements* exist; they're simply not true; nothing answers the descriptions encoded in them. And *that* is the real meaning of "Fred Flintstone/Bart Simpson doesn't exist." The real meaning is clearly not some absurd claim to the effect that there exists some non-existent thing.

(E) there exists some object x such that x cannot exist and therefore does not exist.

2.0 Necessity does not come in degrees

Here is a view about modality that is false, as we'll see in a moment, but is nonetheless a useful starting point[344]:

> (DN) There are different degrees of necessity. "Squares have four sides." This is necessary in the strongest sense. The very idea of a square's failing to have four sides is an incoherent one.
>
> "Metal expands when heated." This is not *as* necessary as "squares

have fours sides," since the idea of a universe in which metal didn't expand when heated is not incoherent. But, given that it describes a law of physics, and is therefore utterly unalterable, it clearly has a high degree of necessity.

"After surgery, people are unusually prone to get infections and must be given antibiotics to stave them off." There is clearly *some* sense in which this is necessarily true. An immunologically normal person who *wasn't* more prone to get infections after surgery is a medical impossibility. And, unless he qualified his point very heavily, a professor of medicine who denied this would be misleading his students. But this statement clearly isn't *as* necessary as the others.

DN is false. There are only two degrees of necessity: having it completely and lacking it completely. When we say that P is "more necessary" than Q, we're saying something about the conditions that must hold for those statements to be true; we're saying that the conditions that must hold for the first to be true are weaker than those that must hold if the second is to be true. "Squares have four sides" is maximally necessary, because it holds *unconditionally.* In any universe at all, no matter how mass-energy is distributed in it, that statement is true.[345] So *no* condition has to be met for that statement to be true. Or, to think of it another way, the only conditions that have to be met are maximally weak: they could be thought of as limiting cases of conditions that are trivially satisfied by any distribution of mass-energy. And to say that "metal expands when heated" is "less necessary" than "squares have four sides" is to say that the former could to be true in a wider class of circumstances than the latter. And, obviously, to say that "after surgery, people are unusually prone to get infections" is "less necessary" than "metal expands when heated" is to say that the former could fail in a wider class of circumstances than the latter. So there aren't really degrees of necessity. Something is either necessary or it isn't. But some statements (e.g., "two is bigger than one") need less help from others to be true—those are the "highly necessary" ones. Others (e.g., "people who wear hats are bad drivers") need a little more help—those are the "less necessary" ones.

3.0 No necessary a posteriori

In 1969, Saul Kripke gave a now famous series of lectures in which he argued that there are non-analytic necessary truths. Here, in my words, is Kripke's argument ("KA" for short):

> (KA) The ancients referred to the last celestial body (besides the Sun) to disappear from the morning sky as "Phosphorous" (or, to be precise, they referred to it with an expression whose Anglicization is "Phosphorous"), and they referred to the first celestial body to appear in the evening sky as "Hesperus" (same qualification). The former is Venus and so is the latter, and the former is therefore *identical* with the latter. The ancients didn't yet know this; and they went for a very long time having no idea that some *one* celestial body appeared in the evening sky before any other and disappeared from the morning sky after any other. But eventually they found this out. What they found out isn't trivial, of course; it isn't in the same category as the fact that Hesperus is identical with Hesperus. What they discovered is a non-trivial, empirical fact.
>
> Henceforth, "FES" and "LMS" will be our abbreviations for the predicates "first celestial body to appear in the evening sky" will be abbreviated as "FES" and, respectively, "last celestial body (besides the Sun) to disappear from the morning sky."
>
> Contrary to what it is extremely tempting to believe:
>
> (HIP[346]) "Hesperus is Phosphorous"
>
> does *not* have the same meaning as
>
> (S*) "the LMS is identical with the FES."
>
> It will take a moment to make it clear why this is so.
>
> Let P be the proposition expressed by HIP. As we just saw, empirical investigation was needed to establish the truth of P; so P is an empirical truth and is therefore non-analytic. It's important to keep the distinction between HIP and P in mind. The ancients didn't have any feelings about the English sentence "Hesperus is identical with Hesperus." And, in making the astronomical discovery that we've

been discussing, what they found out had nothing to do with HIP—or any other sentence. It was an astronomical, not a linguistic, discovery. This point is crucial.

Moving on: If HIP has the same meaning as S*, then P is identical with the proposition meant by S*. And the latter is equivalent with the proposition that:

(P*) exactly *one* thing is an LMS and exactly one thing is an FES, and nothing that has the one property lacks the other. [347]

And the same is true of P, supposing that S* and HIP express the same proposition.

In light of these points, consider a world W* satisfying the following conditions. W* is *semantically* just like our world. (Two worlds or situations are "semantically identical" if the rules that govern word-usage in the one coincide with those that govern it in the other.) So, in W*, people speak English. But, in W*, *Venus* isn't the LMS or the FES. In W*, because of some cataclysm that didn't occur in our world, Mars is the LMS, and Mercury is the FES. So, in W*, "the LMS" picks out Mars, and "the FES" picks out Mercury. This follows from the fact that W* is semantically just like our world. (If John McCain had become President in 2008 instead of Barack Obama, "the U.S. President" would, given the semantic rules that we use, refer to McCain. But, given those very same rules, "Barack Obama" would still refer to Barack Obama. In other words, if x is the person who *in our world* is referred to as "Barack Obama," "Barack Obama" would refer to x in W*. Similarly, if Mars were the LMS, as opposed to Venus, "the LMS" would, given the semantic rules that we use, refer to Mars. But given those same rules, "Venus" would still refer to Venus. In other words, if x is the planet that *n i our world* is referred to as "Venus," "Venus" would refer to x in W*.

"Venus," like "Barack Obama," is thus a *rigid designator.* If x is what "Venus" *in fact* refers to, x is what "Venus" would refer to in *any* world semantically like ours. But "the LMS," like "the current U.S. President," is *not* a rigid designator; it refers to different things in semantically identical situations. (It used to refer to Carter; now it

refers to Obama. But the semantic rules governing that expression haven't changed in the least in the mean time.) In general, proper names are rigid designators. "Hesperus" and "Phosphorous" are proper names. Therefore, *unlike* "the LMS" and "the FES," they are rigid designators. So as long as the semantic rules of English haven't changed, there is some one object x such that (i) each refers to x; and, therefore, such that (ii) each of ‹Hesperus has phi› and ‹Phosphorous has phi› is true exactly if x has phi, a corollary being that (iii) "Hesperus is Phosphorous" (HIP) is true exactly if x is identical with x.

It's obvious that, for any object x, the proposition *x is identical with x* is necessarily true. But it's equally obvious that the proposition expressed by HIP is empirical and, therefore, non-analytic. Therefore, HIP expresses a non-analytic *necessary* truth.

3.1 Kripke's argument evaluated

This argument involves a massive non-sequitur, which the following story exposes. I meet a man. We shake hands. I obviously see what he looks like. He tells me that his name is "Hesperus." (Without worrying about the details, let's assume that I know him to be telling me the truth.) In this context, there is some individual x such that I am learning that:

(HX) "Hesperus" is x's name.

But *what* I am learning must be distinguished from the information *through which I am learning it.* The information relayed to me by my senses isn't confined to HX. It would be absurd to think otherwise. That information is more along the lines of:

(BHX) There is some individual x such that x is standing right in front of me (in my office in Richmond, VA) at this moment (3:00 P.M., March 13, 2009) and such that x has such and such characteristics (he has a mustache, he's over 6-feet tall, etc.) and such that I am being told, by x himself, that "Hesperus" is x's name.

So there is some x such that "Hesperus" merely labels x. "Hesperus" doesn't have for any part of its meaning anything having to do with the

property of being over 6 feet, having a mustache, being in this or that place at this or that time. But the information *through which* I learn whom "Hesperus" labels *does* concern such properties. That information is descriptive; a certain person x is visually represented, and thus described, to me as having certain characteristics (over 6-feet tall, etc.); and, with the help of this information, it is also being made clear to me that "Hesperus" names x. But that descriptive information is not *itself* a part of the semantic rule that, with its help, I am learning—that is to say, it is not itself any part of HX. HX is extremely threadbare; it merely puts a label on a certain object. So the just-mentioned descriptive information has no place in the *semantics* of "Hesperus"; its relevance is confined to the pre-semantic act of assigning it to the right object. That information isn't semantic; it's *pre*-semantic.

The following evening, I receive a phone call. I don't recognize the caller's voice. The caller is obviously using a voice-modulation device, giving it a metallic, threatening sound. And threaten is just what he does; for he says "if you don't move out of town, I'm going to kill you. By the way, my name is Phosphorous." I look at my caller ID; and, sure enough, the call is from one "Phosphorous."

In this context, there is some individual y such that I am learning that:

(PX) "Phosphorous" is y's name.

PX merely registers the fact that a certain name is associated with a certain object. It doesn't say anything (else) *about* that object. It doesn't say whether that object is tall or short, friendly or unfriendly. PX is exceedingly threadbare. But the information *through* which I am learning PX is very different from PX itself. *That* information—that *pre*-semantic information—is along the lines of:

(BPX) There is some individual y such that, a moment ago (7:00 P.M., March 14, 2009), y rang me; such that, using a voice-modulation device to disguise his voice, y is making threatening statements; and such that I am being told, by y himself, that "Phosphorous" is y's name.

Here must we repeat (*mutatis mutandis*) what we said a moment ago. There is some y such that "Phosphorous" merely labels y. "Hesperus" doesn't have for any part of its meaning anything having to do with the property of

talking on the phone with me at a certain time, or of using a voice-modulation device, or of making threats. But the information *through which* I learn who "Phosphorous" labels *does* concern such properties. That information is descriptive; a certain person y is perceptually represented, and thus described, to me as having certain characteristics (calling me at a certain time, making certain statements, etc.) and, with the help of this information, it is also being made clear to me that "Phosphorous" names x. But that descriptive information is not *itself* a part of the semantic rule that, with its help, I am learning—that is to say, it is not itself any part of PX. PX is extremely threadbare; it merely puts a label on a certain object. So the just-mentioned descriptive information has no place in the *semantics* of "Phosphorous," its relevance is confined to the *pre*-semantic act of assigning it to the right object. That information is entirely *pre*-semantic.

But such *pre*-semantic information can have an incalculably profound effect on what is (non-semantically; i.e., non-literally) *conveyed* by utterances. A continuation of our story makes this clear. Right after getting the upsetting phone call, the police call. For some reason, they know about the phone call. They say to me:

(A_1) "Phosphorous is a professor of political science."

There is some y such that A_1 does nothing more than affirm the innocuous proposition that:

(A_2) y is a professor of political science.

But under the circumstances, what will be conveyed to me is some much richer proposition along the lines of:

(A_3) There is some individual y such that, a little while ago (at around 7:00 P.M., March 14, 2009), y rang me; such that, using a voice-modulation device to disguise his voice, y is making threatening statements; such that I am being told, by y himself, that "Phosphorous" is y's name; and, finally, such that y is a professor of political science.

The pre-semantic information embedded in BPX affects what I *take away* from the policeman's utterance of L_1, even though it is no part of its literal meaning. For exactly similar reasons, the pre-semantic information embedded

in BHX affects what I would *take away* from an utterance of:

(B_1) "Hesperus is an avid tennis player."

There is some x such that B_1 does nothing more than affirm the descriptively impoverished proposition that:

(B_2) x is an avid tennis player.

But under the circumstances, what that utterance conveys to me is some much richer proposition along the lines of:

(B_3) There is some individual x such, a couple of days ago (March 13, 2009) at about 3:00 P.M., x is standing right in front of me (in my office in Richmond, VA); such that x has such and such characteristics (he has a mustache, he's over 6-feet tall, etc.); such that "Hesperus" is x's name; and such that x is a political science professor.

We need only add one finishing touch to our story to see what's wrong with Kripke's reasoning. The police arrest Phosphorous. They ask me to come down to the station to talk to them. They point to Phosphorous (who is handcuffed) and say: "that's Phosphorous." Phosphorous, I now see, is none other than Hesperus. The man I see is the mustachioed man I met a few days earlier. The police then say: "Phosphorous also goes by the name of Hesperus. His legal name is 'Hesperus Phosphorous.' But he tends to introduce himself as either 'Hesperus' or 'Phosphorous.' In any case, *Hesperus is Phosphorous.*" The italicized sentence is none other than HIP, of course.

Given what we've already learned, it's obvious that there is some *one* individual z such that this utterance of HIP is true exactly if:

(K) z is identical with z.

Any proposition of that form is trivial. It is always trivial to say of an object that it is identical with itself. But by obvious extensions of what we said a moment ago, when I hear the policeman's utterance of HIP, what I *take* away from it—what it (non-literally) *conveys* to me, given the pre-semantic information at my disposal—won't be at all trivial, as it will be along the lines of:

(K_2) There is some individual x such, a couple of days ago (March 13, 2009) at about 3:00 P.M., x is standing right in front of me (in my office in Richmond, VA; such that x has such and such characteristics (he has a mustache, he's over 6-feet tall, etc.); such that "Hesperus" is x's name; and such that x is a political science professor; and there is some individual y such that, a few hours ago (at around 7:00 P.M., March 14, 2009), y rang me; such that, using a voice-modulation device to disguise his voice, y is making threatening statements; such that I was told by y himself that "Phosphorous" is y's name; and such that y is a professor of political science; moreover, I have just learned from the police that x *is* y.

K_2 is an empirical proposition; it isn't analytic; a fortiori it isn't trivial. *But it's also a contingent truth.* K, on the other hand, *is* a necessary truth. *But it's also analytic.* So we don't have *anything* that is both necessary *and* non-analytic.

Kripke and, after him, many others have professed to find many examples of non-analytic necessary truths; but in each case, an exact analogue of the argument just given shows a confusion similar to the one just identified to be at work. In each case, semantics is being confused with pre-semantics. A defense of this claim is provided in Chapter 8.

4.0 Propositional necessity as microstructural truth: some preliminary concepts

The main modal concepts are necessity, possibility, impossibility, and contingency. All of these can be understood in terms of the concept of necessity. P is possible iff not-P isn't necessary. P is impossible iff not-P is necessary. P is contingent iff neither P nor not-P is necessary.

Whether all modal concepts can be understood in terms of this one concept depends on how inclusive one's conception of modality is. Like Bas van Fraassen (1989), I believe that the property of being probably true is a modal property—that, indeed, for any number n, the property of being probable to degree n is a modal property. It's an open question whether the concept of probable truth can be understood in terms of that necessary truth. I'm inclined to think that it can be, given how natural it is to think of

necessity as a limiting case of probable truth. But until this is resolved, it can't be said that *all* modal notions can be understood in terms of the concept of probability. But there's no denying that necessity is the most important modal concept.

But what *is* necessity? One plausible answer is this: necessary truth is analytic truth; necessity is analyticity. "Squares have four sides" is "analytic," or "analytically true." Indeed, *all* analytically true statements are necessarily true. So it's obviously worth looking into the idea necessity *is* analyticity. Many deny this (for reasons that we'll discuss in a moment); but many, including the present author, hold it. But different authors use the word "analytic" to mean different things, and we must make it clear how we will be using it before we can evaluate the proposal that necessary truth is analytic truth.

Sometimes that word denotes a property of sentences; other times, it denotes a proposition of propositions. A *sentence* is analytic if, given only what it means, it must be true. So "squares have four sides" is analytic. Given *only* what it means, it couldn't possibly be false. A *proposition* is analytic if it's non-empirically true. So the *proposition* that squares have four sides is analytic. No conceivable state of affairs could make it false. We will use the word "analytic" to denote a property of *propositions,* not sentences.

When examining the concept of analyticity, one must make it clear at the outset whether one is talking about sentences or propositions. Otherwise one won't be able to produce a coherent analysis of that concept. This is because there are analytically true *sentences* (or, strictly speaking, sentence-tokens) that have contingent propositions for their meanings. If, at this very moment, I were to say "I am here now," my utterance couldn't fail to be true. Given only the semantic rules that assign truth-conditions to that utterance, it's a logical impossibility that it should be false. But *what it is that I'm saying* could very well be false. Given that I am JMK, that I'm in Richmond, and that it's now March 13, 2009, I am, in producing such an utterance, attributing to JMK the property of being in Richmond, VA, on March 13, 2009. And, quite obviously, it's *not* a necessary truth that I be in that place at that time. Similarly, if I say "I am not here now," my utterance is analytically false: given *only* its semantics—given only what the semantic rules are that assign truths-conditions to it—it's logically impossible that it should be true.

But what it's saying *could* easily have been false; I might not, after all, have been in Richmond, VA, at this very time. It isn't in virtue of what it's saying that a given token of "I am here now" ("I am not here now") must be true (false); it is in virtue of how it is saying it. But it *is* in virtue of what it is saying that a token of "any case of knowledge is a case of true belief" must be true. This shows that there are two very different kinds of sentence-analyticity.

It also shows that sentence-analyticity diverges from proposition-analyticity and, more generally, that sentence-logic diverges from proposition-logic. In other words, the logical relations that propositions bear to one another don't always parallel the logical relation that the corresponding sentences bear to one another. The two sets of interrelations can diverge quite dramatically. (Credit for this astonishing insight goes entirely to P.F. Strawson.) So it's important to bear in mind that, in what follows, the word "analytic" refers to a property of propositions, not sentences. (Later we'll discuss how it is exactly that analytically true (false) sentence-tokens can have contingent truths (falsehoods) for their meanings.)

The word "analytic" is ambiguous in another, equally important respect. As some philosophers use it, "analytic" is synonymous with "trivial" or "tautologous." So "there are 3 feet in a yard" is "analytic" because it's trivial —because, basically, it says nothing. But as other philosophers use that word, analytic truths are often decidedly *non*-trivial—for example, "a class is infinitely large exactly if it can be into a one-one correspondence with a proper subset of itself." When non-trivial truths are described as "analytic," what is meant is that they are non-empirically true. This is what we will mean by it, as earlier stated.

There is one last ambiguity to deal with. As some philosophers use it, the word "analytic" is synonymous with "formally true." Formal truth is a property of *sentences.* It's the property of being "syntactically" true.[348] A sentence is "syntactically true" if any sentence having the same syntactic form is true. So "either John is tall or John is not tall" is formally true, and thus analytic (in one sense of the word), because every sentence having the form "either P or not-P" is true. This obviously isn't how we'll be using the term "analytic," given that, as previously stated, we're using it to pick out a

property of propositions, not sentences. (Propositions don't have syntax. Syntax is a property of complex *expressions*. Propositions aren't expressions.)

A final note: even though sentence-analyticity diverges from proposition-analyticity, the two ultimately coalesce. What does it mean to say that, although the proposition it encodes is synthetic, a given token of "I am here now" is analytic? It means that, the semantic rules of English being what they are, the proposition that any given token of "I am here now" is true is analytic. (Given that any token t_1 of "I" refers to the person who utters t_1, that any token t_2 of "now" refers to the time at which t_2 was uttered, and so on, it follows that a token t of "I am here now" couldn't be false.)

4.1 Propositional necessity as microstructural truth

Analytic propositions are true no matter what. The proposition:

(GG) either grass is green or grass is not green

cannot fail to be true. It doesn't matter what happens or what natural laws govern what happens. Why is this? What is it about GG that makes it incapable of being false? A natural proposal is to say that GG is *structurally* true. In other words, given what its constituents are, along with how those constituents are put together, it cannot fail to be true. All structural truths are analytic truths. Analytic truth is truth in virtue of structure.

This proposal is almost correct, but not quite. Analytic truth is truth in virtue not just of structure but also of *sub*structure. It isn't merely in virtue of its structure that:

(ACK) any case of knowledge is a case of true belief

is analytic; the structures of its *constituents* are involved. Some propositions *structurally* just like ACK are false (e.g., any case of knowledge is a case of false belief). At the same time, there are false propositions that are substructurally just like ACK (e.g., "any case of true belief is a case of knowledge"); and it's obviously because its structure differs from theirs that, unlike those propositions, ACK is analytically true. Analyticity is truth in virtue of structure *along* with substructure. Analytic truth is, as we'll put it,

high-resolution structural truth.

Before we can see what exactly this means, or why it holds, we must make a few non-controversial, but also non-trivial, points about language. Languages are evaluated in terms of how much *expressive power* they have—in other words, in terms of how much can be said in them. The expressive power of a language is to be understood not only in terms of how many propositions can be expressed, but also in terms of how *varied* those propositions are. A language that can express the propositions *snow is white, it is true that snow is white, it is true that it is true that snow is white,* and so on, can express infinitely many propositions. But its expressive resources are minimal; and some other language that can express only a thousand propositions, not a single one of which entails a single one of the others, will grossly exceed the expressive capabilities of the first. And that's why, when assessing the expressive potency of a language, the relevant question isn't so much "how many propositions can it express?" as it is "how much redundancy is there within the set of propositions that it can express?"

These points may be obvious, but they have a non-obvious consequence, namely: as a rule, sentential structure mirrors propositional structure; propositions must have structures similar to those of the sentences that represent them. This doesn't mean that each proposition's structure is perfectly represented by those of all the sentences that express it. (In fact, this is impossible, given that sentences with different structures can express the same proposition.) It doesn't even mean that any proposition's structure is perfectly represented by that of *any* of the sentences that express it. Here's what it does mean: to the extent that the sentences of a given language *fail* to mirror the structures of the propositions they express, that language is expressively frozen. So if L and L* are any two languages, then *ceteris paribus* L's expressive power exceeds L*s to the extent that the structures of L-sentences do a better job than those of L*-sentences of mirroring the structures of the propositions they represent.

Here's the idea. Let L* be a language that has only simple symbols. So the L*-sentence for "snow is colder than water" is some single, syntactically simple symbol, the same being true of every other L*-sentence. Obviously, there's no limit to what L* can express, since there's no limit in principle to

how many primitive symbols can belong to a language. But, since each symbol belonging to L* is simple, it's necessary to *add* a new symbol to L* every time one wants to make a point that isn't the literal meaning of one of its simple expressions. Nothing could be said in L* that hadn't been said before: it would be impossible to express *discoveries.* Symbols would have to be added *ad hoc* to say anything new. But given a language L that has many ways of combining its primitive symbols into complex ones, it's expressive capabilities will dwarf those of L*, so long as the number of primitive expressions belonging to L rises above a certain, extremely small minimum number. L* is totally inflexible. L is highly flexible, and can do a lot with a little. Other things being equal, the extent to which one language's expressive resources exceed those of another is directly proportional to the extent by which sentences of the one language exceed their counterparts in the other in respect of syntactic complexity.

There is only one viable explanation of this last fact: Syntactic structure mirrors propositional structure. The syntactic devices that endow a language with expressive flexibility correspond to the relations by which propositions are constructed out of their constituents and, in particular, by which propositions are constructed out of other propositions. Thus, we can confidently say that the (syntactic) structures of sentences tend to mirror those of the propositions they express.

4.2 Necessity as high-resolution structural truth

We can now say what it means to say that analytic truth is high-resolution structural truth. Given the reasoning just presented, the sentence "either grass is green or grass is not green" must have a structure similar to GG. But, by that same reasoning, GG's structure probably differs from the structure of:

(GA) grass is green and grass is not green.

But GA is not analytically true. (It's analytically false.) Given that GA is false and has the same structure as GG, it cannot possibly be solely in virtue of its structure that GG is true. But we don't just want to give up on the idea that analytic truth is structural truth. An analytically true proposition is *ipso facto* true no matter what contingencies obtain. This entails, or strongly suggests, that it is always in virtue of something *inherent* in it that a

proposition is analytically true, and it's hard to see how the relevant property could fail to be structural in nature. It's obvious that, in the case of GA, what's causing the trouble is that the concept of conjunction (the concept of *and*) is occupying the place that, in GG, is occupied by the concept of disjunction (the concept of *or*). Taken together, these points suggest that it is in virtue of its structure *along with* the structure of one of its constituents—viz. the concept of conjunction—that GG is analytically true.

So it's not quite right to say that analytic truth is structural truth. At the same time, it isn't quite right to identify analytic truth with *sub*structural truth. The proposition:

(NG) it's not the case that grass is green or grass is green

is equivalent with the proposition that grass is not green. Thus, NG is contingently false. But it has the same constituents as the analytically true GG, and is thus substructurally identical with it, even though it's structurally very different.

This suggests that for a proposition to be analytically true is for it to be true in virtue of its structural *and* its substructural properties. In other words, given a proposition's structure along with the structures of its constituents, it's settled whether or not that proposition is analytically true. Thus, a *high-resolution* representation of a proposition's structure settles whether or not it's analytic. (A "low-resolution" representation would be one that failed to disclose substructure.) Thus, *analytic truth is high-resolution structural truth*—and so, therefore, is necessary truth, given that it's identical with analytic truth.

5.0 Quine's denial of informal analytic truth: preliminary concepts

In this section, all references to "analytic" truth are to sentential analytic truths; we will be discussing analytically true sentences, not analytically true propositions.

W.V.O. Quine (1951) says that, with a few trivial exceptions, there is no such thing as sentential analytic truth. To understand exactly what it is that he holds, and why he holds it, we must first make some background points.

Consider the following sentence:

(ST) “if Smith is tall, then it is not the case that Smith is not tall”

Obviously you can’t replace “Smith” with “therefore” or “red.” The result would be syntactically ill-formed nonsense. Thus, such replacements would not be “grammatically admissible.” But you *can* replace “Smith” with “Jones” and “that person.” Such replacements *would* be grammatically admissible, that being why:

(JT) “if Jones is tall, then it is not the case that Jones is not tall”

is a perfectly grammatical sentence.

JT is what results when the occurrences of “Smith” in it are made *uniformly*. In other words, JT is what results when, not just one, but *all* of those occurrences are replaced with the same thing. The following sentence, by contrast, is what results when only *one* of the occurrences of “Smith” in ST is replaced with “Jones” (“NU” is short for “not uniform”):

(NU) “if Jones is tall, then it is not the case that Jenkins is not tall.”

Thus, a syntactically well-formed sentence S cannot be false if it is the result of uniformly replacing the occurrences in ST of “Smith” (or, for reasons analogous to those just given, of “tall”) with some other expression.

In general, a sentence S is *formally* true if there is some expression E occurring in S such that no syntactically well-formed sentence S* is false if S* is what results when the occurrences in S of E are uniformly replaced with occurrences of some other expression.

It’s generally held—and the present author himself holds this—that there are *informal* analytic truths. Consider, for example, the following sentence (“JTB” is short for “justified true belief”):

(JTB) “one can’t know something without believing it.”

JTB is analytically true. Given *only* a knowledge of what it means, one knows (or is in a position to know) that it is true. It isn’t like “there are nine planets,” which, if true, can be established only on the basis of experimental data. No—JTB is *inherently* true. The denial of it (“one *can* know something without believing it”) is prevented, by its very own meaning, from being true; and it (JTB) is *required* by its very own meaning to be true.

But JTB isn't *formally* true. If you replace "know" with "doubt," the result is a false sentence. In fact, given *any* expression in JTB, replacing it (in a uniform and grammatically admissible manner) with some other expression results in falsehood. For example, the sentence "one can't deny something without affirming it" is a false sentence that results from replacements of the sort just described.

It seems pretty clear, then, that there are informal analytic truths. And this is what most analytic philosophers hold, including the present author. But Quine denies it. Quine thinks that, apart from formal analytic truths, there are *no* analytic truths. This thesis of his is famous, and so is his argument for it.

5.1 Quine's denial of informal analytic truth

Here is Quine's argument. Consider the sentence:

(B) Bachelors are unmarried men.

If any sentence is analytic, (B) is one of them. According to Quine, if (B) is analytic, that is in virtue of the fact that "bachelor" and "unmarried man" are synonyms. Quine says that *whenever* a sentence is analytic, that is because it holds entirely in virtue of some synonymy relation.

Quine then asks the following question: What is it for two expressions to be synonyms? Quine gives the following (correct) response: It is for them to be such that inter-substitutions of them preserve meaning. "Foe" and "enemy" are synonymous because you cannot change sentence meaning by replacing "foe" with "enemy." Put another way, "foe" and "enemy" mean the same thing because, if a sentence S contains an occurrence of "foe," and that occurrence is replaced with an occurrence of "enemy," the resulting sentence has the same meaning as S.

Let me explain what this last sentence means. Consider the sentence: "Smith is a foe." Now replace the word "foe" with "enemy." What results is "Smith is an enemy." The sentence "Smith is an enemy" has precisely the same meaning as "Smith is a foe."

Let S and S* be *any* two sentences that are exactly alike *except* for the fact that, in the place where S contains the word "foe," S* contains the word

"enemy." What we just said about "Smith is an enemy" and "Smith is a foe" is true of S and S*. And that is why "foe" and "enemy" are synonyms. Inter-substituting synonyms doesn't change sentence-meaning—and for two expressions to be synonymous just *is* for them to be such that inter-substituting them doesn't ever change sentence-meaning.

But, Quine asks, what is it for two sentences to have the same "meaning"? What is it for "Smith is an enemy" and "Smith is a foe" to mean the same thing? According to Quine, for two statements to have the same meaning is for them to be *analytically equivalent.* In other words, it is for them to entail, and be entailed by, exactly the same statements.

Let me illustrate this last point. "Smith is an enemy" entails "Smith is somebody who is attempting to do me ill," and so does "Smith is a foe." The statement "Smith is somebody whose sole objective in life is to make my life as unpleasant as possible" entails "Smith is an enemy," and it *also* entails "Smith is a foe." Given *any* statement S that entails "Smith is an enemy," S *also* entails "Smith is a foe"; and given *any* statement S* such that "Smith is an enemy" entails S*, "Smith is a foe" *also* entails S*. Thus, "Smith is an enemy" and "Smith is a foe" entail, and are entailed by, exactly the same statements—and that, according to Quine, is what it is for them to have the same meaning.

But notice that we have gone in a circle. We defined analyticity in terms of synonymy. A recap will make this clear:

1. (B) is analytic because "bachelor" and "unmarried man" are synonyms: in general, if a sentence is analytic, that is because it holds true entirely in virtue of some synonymy relation.

We then defined synonymy in terms of sameness of meaning:

2. For two expressions to be synonymous is for them to be such that inter-substitutions of them preserve sentence meaning.

But we then defined sameness of meaning in terms of analyticity:

3. For two statements to have the same meaning is for them to be *analytically equivalent.* In other words, it is for them to entail, and be entailed by, exactly the same statements.

1 defines analyticity in terms of 2 and, thus, in terms of synonymy. 2 defines synonymy in terms of 3, and thus in terms of analyticity. Thus, analyticity has been defined in terms of itself. We have gone in a vicious circle. Thus, Quine concludes, the concept of analyticity is viciously circular and therefore incoherent and, consequently, there is no such thing as analytic truth: all truths are synthetic.

5.2 Evaluating Quine's argument

First of all, Quine is wrong to hold that analyticity is to be understood in terms of synonymy. Consider the sentence:

(T) The unique number n such that n = 2 if the interior angles of a Euclidean triangle add up to 180° and n = 3 otherwise is identical with the unique number m such that 9^m = 90 minus 9.

(T) is analytic. It isn't empirical. No empirical knowledge, over and above such as is needed to know what meaning it bears, is required to know whether it's true. So it's true in virtue of what it means and is thus analytic. (To be sure, unlike some analytic sentences, it isn't *formally* true. But since the very thing that Quine is trying to show is that there is no informal analytic truth, he cannot, without assuming the truth of the very thing he's trying to show, claim on those grounds that T is not analytic.) So if Quine is right, T holds in virtue of some synonymy relation. What, in this case, would the relevant synonyms be? Presumably they would be:

(*) "the unique number n such that n = 2 if the interior angles of a Euclidean triangle add up to 180° and n = 3 otherwise"

and

(#) "the unique number m such that 9^m = 90 minus 9."

But, even though they *refer* to the same thing (namely, the number two), * and # obviously don't have the same *meaning;* and Quine is therefore wrong to say that, if there are analytic truths, they always hold in virtue of synonymy relations.

5.3 Analytic truth non-circularly definable as non-

empirical truth

Quine's larger point seems to be that there is no non-circular definition of the term "analytic truth." But that is false: a sentence is analytic if (a) it is true and (b) it does not hold in virtue of *empirical* facts. Put simply: analytic truth is non-empirical truth. "Circles are figures of uniform curvature" is true and, moreover, it holds, not in virtue of empirical facts (i.e., not in virtue of how objects are *in fact* shaped), but in virtue of facts about the structures of the properties *circle, curvature,* and so on. Of course, being a nominalist, Quine would reject the idea that there are non-empirical truths. (After all, a nominalist believes that everything is spatiotemporal: there are no platonic entities and thus no properties.) But in attempting to establish that all truths are synthetic, Quine is, in effect, trying to *establish* that there are no empirical truths. For an empirical truth is, by definition, one that can be learned only through observation and therefore cannot be learned through any non-sensory modality such as conceptual analysis. So Quine would be guilty of begging the question were he to say that, since all truth is empirical, it is not an option to define "analytic" as "non-empirically true."

5.4 Quine's position self-defeating

Also, the view that there are no analytic statements is self-defeating.[349] The statement:

(^) There are no empirical statements

is equivalent to the statement:

(^^) for all values of S, the statement *nothing follows from S* follows from the statement *S is a statement.*

In other words, if you say that there are no analytic statements, you are saying that nothing follows from any statement. But if you say *that,* then you are saying that the statement *nothing follows from S* follows from the statement *S is a statement.* Thus, the view that there is no analytic truth entails that *there is* analytic truth, and is therefore self-refuting.

5.5 Why Quine's position entails that nothing can

confirm anything

There is another reason to hold that, contrary to what Quine says, there are in fact analytic truths. To say that there is no analytic truth is to say that all truths are empirical truths. For the sake of argument, let us suppose that all truths are indeed empirical. In that case, given any two statements S_1 and S_2, it would be an *empirical* question—one to be decided observational grounds—whether S_1 supported S_2; and, supposing that S_1 did support S_2, it would be an empirical question to what degree it did so.

This means that, for any values m_1 and m_2, and for any observations $A_1 \ldots A_n$, it would be an empirical question whether or not:

(S_3) Observations $A_1 \ldots A_n$ confirm to degree m1 that S_1 confirms S_2 to degree m_2,

were correct. And it means that, for any value m_3, and for any observations $B_1 \ldots B_o$, it would be an empirical question whether or not:

(S_4) observations $B_1 \ldots B_o$, confirm to degree m_3 that observations $A_1 \ldots A_n$ confirm to degree m1 that S_1 confirms S_2 to degree m_2,

were correct. And so on *ad infinitum.* Thus, the supposition that all truth is empirical is viciously regressive. Thus, if there is no analytic truth—that is, if all truth were empirical, then nothing could confirm anything and, as Laurence Bonjour (1998) puts it, all inquiry would "grind to a halt."

6.0 Why there are necessarily true propositions

First a terminological point: If a proposition is true, but it isn't necessarily true, it is "contingently true," or just "contingent"; and if a proposition is false, but it isn't necessarily false, it is "contingently false."

It seems clear that *some* truths are contingent. Gore *might* have become President in 2001. He didn't. Bush became President instead. But Bush didn't *have* to become President. So it's a contingent truth that Bush became President. In general, things don't *have* to be the way they are; they could be different. In any case, this is what we think, and it's what we'll assume in this

context.

Given this assumption, we must ask the following question: Are any propositions necessarily true? Yes. Why is this? Because the assumption that no propositions are necessarily true is self-undermining. Let me now explain why.

For argument's sake, suppose that:

(NN) No propositions are necessarily true.

First of all, is NN itself necessarily true or not? It has to be one or the other. If NN is a necessary truth, it's a counterexample to itself and is therefore false. So let's suppose for argument's sake that NN is contingently true. In that case, it *could* be false. In other words, there is some possible world W where some proposition P is necessarily true. A proposition is necessarily true iff it's true in all possible worlds. So given that P is necessarily true in W, it follows that P is true in *every* world. Thus, P is true in our world, and P is true in every world. Thus, P is *necessarily* true in our world, and NN is therefore false. To sum up: NN is self-defeating if it's necessarily true, and NN is self-defeating if it's true but not necessarily so. Therefore, NN is false.

A similar argument: if NN is true, then the statement (i) "P isn't necessarily true" necessarily follows the statement (ii) "P is a proposition," and it is therefore a necessary *truth* that (i) isn't false if (ii) is true. So NN cannot be true unless it's false, and it's therefore false.

6.1 Knowledge of necessary truth a prerequisite for knowledge of contingent truth

Nonetheless, some philosophers have denied that there are necessary propositions. We just considered Quine's argument for this. Given how contrived that argument is, how lacking in logical integrity and devoid of intuitive force, it's hard to believe that Quine's rejection of necessary truth was based entirely on it. Quine, one would suspect, must have had some pre-existing agenda and he constructed that argument in the hopes that it would legitimate that agenda.

Such suspicions would be correct. Throughout his long career, Quine was

unrelenting in his advocacy of a particularly extreme form of empiricism.

To understand Quine's empiricism, we must note a fact about empiricism in general. Empiricism says that you know only what your senses are telling you. Unless you assume the truth of principles that cannot be learned empirically, you can't show that your perceptions correspond to any transperceptual reality. For this reason, empiricism tends to collapse into idealism, this being why Berkeley, who has a hardcore empiricist, was also an idealist.

Quine's empiricism is in some ways antithetical to Berkeley's empiricism. Berkeley was led by his empiricism to deny the non-mental. Quine was led by his empiricism to deny the mental. According to Quine, there is no observational evidence of mental states. One sees bodies. One *doesn't* see or otherwise sense-perceive mental states. Quine's position is that mental states either don't exist or are identical with the bodily behaviors that we ordinarily, and wrongly in Quine's view, believe to be *expressions* of mental states. Thus, Quine was a behaviorist. (In Chapter 12, it is shown why behaviorism is false.)

Quine also held that there is no a priori knowledge. There is, he believed, no knowledge that is hardwired into us and that regulates the intake and subsequent processing of sensory information. Quine also believed that our linguistic behavior can be explained on the assumption that, at birth, we are blank slates and that it is entirely through a process of Pavlovian conditioning that we come to use words in the way we do.

This doctrine is *also* known as "behaviorism." Behaviorism in *this* sense is a psychological theory that tries to answer the question "how does a human being learn to speak a language?" Behaviorism in the other sense is a philosophical theory that tries to answer the question "what are mental states?" To avoid confusion, I'll refer to the first doctrine as "linguistic behaviorism."

Linguistic behaviorism is incoherent. To say "that's a fire" and *mean* it, I must have beliefs as to what that expression means and as to what circumstances would warrant my uttering it. If I reflexively bark it out when I see a fire, I'm no more affirming that there's a fire than I am if I start sweating in response to the heat given off by a nearby fire. These points are

substantiated in Chapter 4.

Also, in 1957, Noam Chomsky published his groundbreaking work *Syntactic Structures,* which, along with the many other important works that Chomsky was soon to publish, made it clear that Quine's linguistic behaviorism simply didn't model the relevant facts. Chomsky proposed an alternative theory, according to which inborn cognitive structures are largely, though obviously not entirely, determinative of how we learn and use language. Unlike Quine's theory, Chomsky's is splendidly borne out by the facts.

Quine and others accused Chomsky of putting forth a priori theories that lacked an adequate empirical grounding. This is both false and confused. It's true that *according* to Chomsky's theory we have a priori knowledge. But it doesn't follow that Chomsky's theory is itself a priori. Chomsky's theory is based on empirical data; and, unlike Quine's theory, it fits that data. And even though, according to Quine's theory, there is no a priori knowledge, it doesn't follow that Quine's theory is itself empirically well-founded. It isn't. It's rooted in the a priori, and easily disconfirmed, presumption that all psychological phenomena are to be explained in the same way as the salivating of Pavlov's dog.

One of the great ironies of empiricism is that it's one of the most unempirical theories of all time. People learn languages. Dogs don't. A dog wouldn't even learn the rudiments of a language were it to experience the sensory simulations that enable a human being to master one. This shows that, unlike dogs, human beings are endowed with innate cognitive structures that enable them to derive an integrated and complex body of knowledge from such stimulations. The same point *mutatis mutandis* holds of any cognitive skill that human beings can acquire that animals cannot. Empiricism is therefore inconsistent with empirical evidence of the most blatant and widespread kind.

In any case, despite the vast and ever-growing body of evidence in support of Chomsky's anti-Quinean views, Quine stuck to his guns, and in the 45 years that passed between the publication of Syntactic Structures and his death, Quine didn't soften his position. Not only did he not concede that Chomsky was at all right; he didn't even concede that mental entities even

exist, except to the (non-existent) extent to which they can be identified with overt behaviors.[350]

In any case, it's clear that Quine's rejection of analytic truth wasn't the effect, but was instead the cause, of his acceptance of the bizarre and unwieldy argument of his that we considered earlier. Quine obviously produced that argument *ad hoc*, hoping to give his idiosyncratic form of empiricism a solid foundation. Quine's real reason for denying the existence of analytic truth lay in his staunch empiricism. All philosophers, and indeed non-philosophers, who deny the existence of analytic truth do so because they are staunch empiricists.

Thus, when authors deny that there are necessary truths, the underlying sentiment, if not the immediately operative rationale, is this:

> (NT) If they exist, necessary truths hold no matter what. They hold no matter how the world is and, therefore, no matter what observations one has. So, if they exist, they can be known *independently* of observation. But it's unscientific to think that anything could be known in any way other than through empirical observations. Therefore, there are no necessary truths.

6.2 Why NT is a bad argument

All knowledge of the spatiotemporal world is rooted in observation. But there's more to doing science than making observations. The observations are only a means to an end, the end being the acquisition of theoretical knowledge. We'll now see that non-observational knowledge is a precondition for theoretical knowledge and also that it is always necessary truths that are known non-observationally.

Our senses give us raw data. Raw data is the foundation of all empirical knowledge. But data cannot interpret itself. And even if it could, it wouldn't relieve of you of the burden of having to interpret the data for yourself. For argument's sake, suppose that the data were self-interpreting, and that its self-interpretations were disclosed to you in periodic reports. Any such report would *itself* be given to you in terms of raw data that you'd have to analyze; so you couldn't even understand the report unless you already knew how to

interpret the data. Second, once you did understand the report, you'd have no good reason to trust it unless you could find some independent corroboration for it. Finding such corroboration would involve coming up with your own interpretation of the data and seeing how well it matched the interpretation in the report. So even if *per impossible* perceptual data could self-interpret, one could learn nothing about the external world from those interpretations unless one could interpret the data for oneself. It is, therefore, totally incoherent to suppose that one could know on the basis of observation alone how to interpret the data of observation; and since it's totally incoherent to suppose that one could have theoretical or, therefore, scientific knowledge without interpreting the data, it is totally incoherent to suppose that any science should be *strictly* observation-based. There is necessarily a non-observational component to the acquisition of scientific knowledge.

Interpreting data involves being able to discern abstract resemblances between different concrete situations. Doing this, in its turn, involves knowing necessary truths. Seeing that two objects are alike in respect of (e.g.) their shape involves seeing that they fulfill conditions that, if not *fulfilled* by an object, necessarily preclude it from having the shape in question. Seeing that otherwise dissimilar objects x and y are both square-shaped involves knowing that, were an object to have more sides than x or y, it would *for that very reason* fail to have the same shape as them, and it also involves knowing that, if an object had a different color from x or y, it would not *for that very reason* fail to have the same shape as them. One can't know that x and y have the same shape unless one knows that an object's having four sides *necessarily* disqualifies it from being shape-identical with those objects. In general, one cannot discern *any* abstract similarities between *any* two concrete situations unless one knows, for at least some property, what conditions *must* be fulfilled if a given object is to have that property. So knowledge of necessary truths is a precondition for being able to recognize resemblances between things. Since knowledge of such resemblances is the first, and the most rudimentary phase, of scientific endeavor, it is radically incoherent to suppose that there should be any scientific knowledge without knowledge of at least some necessary truths.

W.V.O. Quine tried to provide a rigorous proof of NT. We considered this proof in Sections 5.0–5.5 of the present chapter and found it to be

wanting. But notice that the very idea of trying to *prove* that contention is a self-defeating one. For there are no proofs if all truth is contingent, since proving something is showing that it *must* be the case. (More accurately, to prove conclusion C on the basis of premises P is to show it to be necessarily true that if P, then C.) Thus, to prove NN to be true is to prove it to be false. So a certain incoherence is embedded in the very idea of trying to show that there are no necessary statements.

7.0 Analytic truth vs. formal truth

Formal truth isn't analytic truth. We'll discuss this in this section, and we'll also see what exactly it means to describe a truth as "formal." A statement is analytically true if its negation is incoherent. "Squares have four sides" is analytic. Why? Because "squares don't have four sides" is incoherent, and so is "anything that is literate is animate." The reason being that it makes no sense to suppose that squares might fail to have four sides or that literate beings might fail to be animate.

"Right now, in 2009, Barack Obama is the U.S. President" is a non-analytic truth. This is because "Barack Obama is not President in 2009," though false, makes sense. If that sentence didn't make sense, then in 2007 it wouldn't have been reasonable to wonder whether Obama would be President in 2009. But it was.

Non-analytic truths are ***empirical*** truths. An empirical truth is one that can be known only through sensory observation. It is only on the basis of sense-perception that one can know that Barack Obama is currently the U.S. President. That is why "Barack Obama is currently the U.S. President" expresses an empirical truth. But it is entirely through insight into its meaning, and into its *entailment-relations,* that one knows that nothing could be to be identical with itself; that being why that truth is analytic.

In the *Critique of Pure Reason,* Immanuel Kant (1724–1804) infamously claimed that there are non-analytic (or, as he referred to them, *synthetic*) truths that can be learned in a *non-observational manner.* He was wrong, as we'll see in Section 9.2.

7.1 Empirical vs. pseudo-empirical truth

One must distinguish genuine empirical knowledge from analytic knowledge the acquisition of which is *triggered* by empirical knowledge. Consider the principle that:

(DN[351]) if a statement is false, then the negation of its negation is false.

For example, if Smith is in Richmond, then it is false that Smith is not in Richmond. DN is one of the cornerstones of logic.

Learning DN may involve hearing or reading, and therefore seeing, some affirmation of it. But you don't learn DN empirically—that is, it isn't through sensory observation that you learn it. But, if you know DN, it is because you see that, given its meaning, it would make no sense to deny it. Sense-perceptions may well have been needed to *initiate* or *trigger* the reflections on the basis of which you learned DN. But such sense-perceptions were themselves no part of that discovery process. They were part of the *pre*-discovery process—much as turning the ignition in a car is part of the *pre*-journey, and not the journey proper.[352]

Given a true statement S, there is a simple test for determining whether it is empirical or analytic. If, in order to *justify* acceptance of S, you must cite empirical data, it's empirical. To justify your acceptance of the statement that Smith has gained 20 pounds in the last year, you must cite empirical data (e.g., sense-perceptions that you have had of Smith or of affirmations on the part of reliable sources as to Smith's weight situation). But your acceptance of DN can be justified only by insight into its meaning; it cannot possibly be justified on observational grounds. Of course, one might justify one's acceptant of DN by citing the fact that one has heard some authority affirm it; and, of course, one's knowledge of that fact is observation-based. But those observations justify DN *only* to the extent that the authority figures in question understand why, for reasons that lie entirely in DN's content and having nothing to with anything one perceives, that statement must be true. Testimony-based knowledge of analytic truths, if it exists, is parasitic on non-observational knowledge of them.

And it's doubtful whether such knowledge really exists. If I don't understand why DN is true, and I accept it only because somebody I trust does so, then I don't really know it. What *I* accept is not, it seems to me, quite

what the authority figure in question accepts. That person really accepts DN. But what I accept isn't DN, and is instead some statement along the lines of: "since so and so is authoritative, and since so and so affirmed DN, that statement is probably right"—which, unlike DN itself, is empirical.

Sense-perception adds to one's knowledge in two different ways. In some cases, it does so by *transmitting* information about the external world. If I see Jim playing soccer, my visual perception adds to my knowledge by transmitting to me the correct message that Jim is playing soccer. In other cases, sense-perception merely *activates* some purely ratiocinative process. The ratiocination-activating role of sense-perceptions must be distinguished from its information-transmitting role. **Empiricism,** the doctrine that all knowledge is derived from the senses, is based on a failure to make this distinction. See Chapter 13.

7.2 Analytic knowledge of principles vs. empirical knowledge of the fact that certain sentences *express* those principles

It's an empirical fact that the *sentence* "squares have four sides" means what it does. Those sounds could mean anything. They could mean that squares don't have four sides or that everybody owns a Rolls Royce. A brilliant extraterrestrial could reasonably wonder what that sentence means. But such a being could *not* reasonably wonder whether squares have four sides. Thus, it's an *empirical* truth that the *sentence* "squares have four sides" is to the effect that squares have four sides. But it's an analytic, non-empirical fact that squares have four sides.

7.3 Analyticity vs. triviality

The concept of analytic truth is typically *illustrated* in connection with trivial statements (e.g., "any given thing is identical with itself"). But there are ***non-trivial*** analytic truths. Here are some examples:

* Given only that an institution is flourishing, it cannot be inferred that the individuals composing it are flourishing; and given that the individuals composing an institution are flourishing, it cannot be inferred that the

institution itself is flourishing.

*No finite set can be put into a one-one correspondence with a proper subset of itself.

*Any infinite set can be put into a one-one correspondence with a proper subset of itself. *Anything sentient is, at any given point, unaware of at least some of its mental states.

*There are cases of justified true belief that are not knowledge.

*There are continuous functions that cannot be differentiated at any point.

*Two objects that cannot enter into causal relations with each other cannot have spatial or temporal positions relative to each other.

Also, it's an analytic truth, and also a non-trivial one, that there are non-trivial analytic truths. We saw this in Chapters 1 and 7.

7.4 No synthetic non-empirical truths: an outline of an argument

In the *Critique of Pure Reason* Kant claimed that there are non-analytic (or, as he called them, *synthetic*) truths that can be learned in a *non-observational manner.*[353] He was wrong. To say why, we must first define some terms.

The ***negation*** of a statement P is not-P. (The negation of *snow is white* is *snow is not white.*) A statement is ***coherent*** if what it says isn't ruled out solely by considerations of how things *must* be. An analytic statement is one whose negation is incoherent. (It's analytic that triangles have three sides, since a triangle's having four sides can be ruled out entirely on the basis of consideration of how things must be.)

Now for the argument. A statement may be non-analytic either because its negation is analytically true (*triangles don't have four sides* is non-analytic for this reason) or because both it and its negation are coherent (*JMK is in Richmond* is non-analytic for this reason). A statement that satisfies the second condition cannot be known to be false (or true) merely on the basis of how things must be, and can known to be false (or true) only on the basis of how things are. How things are can be known only through observation. Anything so known is *ipso facto* empirical. So any non-analytic truth is

empirical.

A much more developed version of this argument is given in Chapter 18.

7.5 Formal truth

The concept of *formal truth* has an important place in philosophical inquiry. It is sometimes propositions that are described as "formally true." But this is quite rare, and it's almost always expressions that are so described. To the extent that formal truth is a property of expressions, it must be a property of sentence-*tokens,* not of sentence-types, since the latter, as we saw in Chapter 4, are never true or false. And, in what follows, we must bear in mind that, so far as any expressions are formally true, it is sentence-tokens, not sentence-types.

Here are some examples of sentence-types whose tokens would universally be agreed to be "formally true":

(1) If Bill is tall, then it is not the case that Bill is not tall.

(2) If everything has a given property, then any given thing has that property.

(3) If all F's are G's, and no G's are H's, then no F's are H's.

(4) Either Helsinki is in Finland or Helsinki is not in Finland.

We saw in Chapter 4 that sentence-tokens are the only expressions that are true or false—that sentence-types are neither, since their meanings are rules that assign truths and falsehoods to their tokens, and aren't truths or falsehoods per se. But to avoid verbosity, we'll often speak as though sentence-types per se can be formally true. So we'll say, for example, "(1) is formally true," and we'll discuss the conditions that "sentences" must satisfy to be formally true. But it's crucial to bear in mind that this is *only* an expedient way of talking about sentence-*tokens*—of saying that *tokens* of (1) are formally true, etc.

What is "formal truth"? What does it mean to describe (1), for example, as "formally true"? Here's the canonical answer ("FT" stands for "formal truth," of course):

(FT) Formal truth is *syntactic* truth. A sentence is "syntactically true" if, given only its syntactic structure, it cannot fail to be true. Thus, a sentence is

formally true if no sentence having the same syntactic structure is false.

FT cannot be accepted. A linguist would say that (1) “grass is green or grass is not green” has the same syntax as (2) “grass is green and grass is not green.” As linguists use the word “syntax,” two expressions have the same syntax if the way in which the one is built out of its parts is identical with the way in which the other is built out of *its* parts. For example, “Bill hates Ted” and “Laura loves Sally,” though having very different meanings, have the same *syntax,* the reason being that they’re structurally identical. The one is built out of its simple constituents in the very way that the other is built out of *its* constituents. Thus, it isn’t because of any structural differences between them that they have different meanings.

The relevant difference is *sub*structural. It lies in the fact that the *meaning* of “and” is different from that of “or.” So the difference between (1) and (2) is like the difference between “John is tall” and “Mary is short”: it lies entirely in the fact that the primitive, semantically unstructured constituents of the one have different meanings from their counterparts in the other. Thus, relative to what *linguists* mean by “syntax” (and “syntactic,” etc.), it’s a non-starter to say that “formal truth” is syntactic truth. Logicians and philosophers obviously have in mind some very different understanding of the word “syntax” when they identify “formal” truth with “syntactic” truth. They *don’t* have in mind sentential (or phrasal) structure. What *do* they have in mind? What do philosophers mean when they say that a formally true sentence is one that holds entirely in virtue of its “syntax”? What do they mean when they describe some statement as a “formal truth”?

7.5.1 Quine on formal truth

W.V.O. Quine (1908–2000) gave a characteristically intelligent and straightforward answer to this important question: a sentence S is formally true if, given any expression in S that isn’t a *logical constant,* it isn’t possible to produce a false sentence by replacing that expression with another expression of the appropriate grammatical category. A “logical constant” is an expression on the following list: “if . . . then . . . ,” “either . . . or . . . ,” “not,” “for all,” “for some,” “=”—and any obvious derivatives of any of these expressions (e.g., “is identical with,” “given any,” “it is not the case that,”

etc.)[354]

Consider the sentence "grass is green or grass is not green." Replacing "grass" with "fire" or "Bob" results in a truth, as does replacing "green" with "cool" or "voluminous." (It's assumed that the replacements are uniformly made: if a single occurrence of "grass" is replaced with "fire," so is every other such occurrence.) Given any expression in (1) *other* than the logical constants in it (viz. "either . . . or . . ." and "not"), replacing that expression with another belonging to the same grammatical category results in truth.

There are two problems with Quine's analysis. First, it doesn't say what is so special about those items; it doesn't say what these so-called "logical constants" have in common with one another that other expressions *not* on the list *don't* have in common with them. And until that information is provided, there's nothing to prevent one from extending that list in arbitrary ways. Unless that list was generated by some rule, anything can be added to it without diminishing its integrity, since it has no integrity to begin with. But if it was generated by some rule, then Quine's answer fails to identify it and thus fails to say what formal truth is.

An analogy may help. When asked to define the expression "number," I simply *list* the various different kinds of numbers ("real," "whole," "rational," "imaginary," etc.). Have I answered the question? No. Even though I've listed the various different kinds of numbers, I haven't said what numbers are, and I thus haven't provided any *reason* why fractions are on that list but birds are not. Accordingly, my answer doesn't give anyone a reason not to extend that list as he or she pleases—not to decide that birds and rocks are numbers. Of course, nobody *would* extend the list in such an absurd way. But that's only because our intuitions strongly suggest that there is *some* characteristic had in common by reals, rationals, etc., that birds and rocks don't have.

Second, there are truths not covered by Quine's analysis that don't appear to differ in any relevant respect from those covered by that analysis. Many modal statements are in this category—for example, "if it's necessarily the case that Smith is a mammal if Smith is a dog, then if it's necessarily the case that Smith is a mammal if it's necessarily the case that Smith is a dog." There doesn't seem to be any relevant difference between *that* statement and

one that conforms to Quine's analysis—for example, "if Smith is a mammal if Smith is a dog, and Smith is an animal if Smith is mammal, then Smith is an animal if Smith is a dog." So there doesn't seem to be any good reason why "necessarily," unlike "if . . . then . . . ," doesn't make it onto Quine's list.

Of course, one could deal with this by adding "necessarily" to that list. But then we'd have to ask: "what makes *that* extension of the list legitimate?," bringing us back to the criticism made a moment ago.

7.5.2 Why formal truth ≠ topic-neutral truth

Some have argued that formal truth is maximally generic, or "topic-neutral," truth. The idea is that "x = x" is logically true because it's completely general: it doesn't belong to any one discipline and is presupposed by all of them.[355]

But this doesn't work. There are maximally generic truths that aren't formal—for example, "nothing can wholly pre-exist itself," "nothing spatiotemporal can lack causal properties," "nothing non-spatiotemporal can have causal properties," "inductive inference isn't an appropriate method of reasoning for non-empirical disciplines." Given any one of these sentences, it's easy to show that there are false sentences having the same syntactic form. For example, the first of those sentences has the same form as "nothing spatiotemporal can have causal properties," which is obviously false.

One could argue that these truths *aren't* sufficiently generic. But then it would be very unclear what exactly "sufficiently generic" meant in this context, and the suspicion would arise that, so far as it had any coherent meaning, "sufficiently generic" meant "not a counterexample to this particular theory of formal truth."

Also, whether a sentence expresses a generic truth has nothing to do with its form. The truth expressed by "if everything has a given property, then any given thing has that property" is surely a maximally generic one. But that truth could obviously be expressed by a primitive, syntactically unstructured symbol. In fact, there is *no* form that sentences must have if they are to express that truth or any other. So, if "formal truth" is taken to be a characteristic of *expressions* of some kind, then it's simply false to identify "formal truth" with "maximally generic" truth. And if "formal truth" isn't

meant in this way, then it's simply irrelevant to our current inquiry.

7.5.3 An alternative analysis of formal truth

A sentence is formally true if it is an instance of an open-sentence that is true for all values of its free variables. (An "open-sentence" is an expression that contains a free variable, and is therefore neither true nor false, but is otherwise just like a sentence. So "x is tall" is an open-sentence.[356]) For example:

(5) "If Smith is at home, then it's not the case that Smith is not at home"

is a formal truth because it's an instance of the open-sentence "if x is phi, then it's not the case that x is not phi," which is true for all values of x. (We'll see that (5) is an instance of many such open-sentences. But its being an instance of just one is enough for its being formally true.) Given any formal truth, the same holds.

If an open-sentence is true for all values of its variables, we'll say that it's "universally true." So "x = x" is universally true (that being why "Smith = Smith" is "formally" true). It must be kept in mind that this is just an abbreviation. "x = x" is neither true nor false. No open sentence is true or false.

In light of these points, consider the following expressions:

(1) Bill Gates is wealthy.

(2) x is wealthy.

(3) Bill Gates is phi.

(4) x is phi.

(1) is a true sentence. (2)–(4) are neither true nor false. Given (1), (2) is what results when "Bill Gates" is replaced with a variable, (3) is what results when "wealthy" is replaced with a variable, and (4) is what results when both are replaced with variables. (See Chapter 7.) Of course, actual sentences are formed out of these open-sentences by replacing the variables in (6)–(9) with constants. In each case, this procedure may yield either a true sentence or a false one. For example, replacing the variable in (2) with "Bill Gates" yields a truth and replacing it with "Mike Smith," the name of an impecunious

friend of mine, yields a falsehood.

Bearing these points in mind, consider the following expressions:

(5) If Smith is at home, then it's not the case that Smith is not at home.

(6) If x is at home, then it's not the case that x is not at home.

(7) If Smith is phi, then it's not the case that Smith is not phi.

(8) If x is phi, then it's not the case that x is not phi.

(9) If P, then it's not the case that not-P.

Each of (6)–(9) is formed by replacing one or more expressions in (5) with a variable. And the following sentences result when, in each case, the variables are bound by universal quantifiers.

(6UG) For any individual x, if x is at home, then it's not the case that x is not at home.

(7UG) For any property phi, if Smith is phi, then it's not the case that Smith is not phi.

(8UG) For any individual x and any property phi, if x is phi, then it's not the case that x is not phi.

(9UG) For any proposition P, if P, then it's not the case that not-P.

This suggests that a sentence is formally true if it's an instance of a universally true open-sentence.

To take another example:

(10) If it's necessarily the case that Smith is a mammal if Smith is a dog, then it's necessarily the case that Smith is a mammal if it's necessarily the case that Smith is a dog

is formally true, because it's an instance of the following universally correct open-sentences:

(11) If it's necessarily the case that x is a mammal if x is a dog, then it's necessarily the case that x is a mammal if it's necessarily the case that x is a dog.

(12) If it's necessarily the case that Smith is a mammal if P, then it's necessarily the case that Smith is a mammal if it's necessarily the case

that P.

(13) If it's necessarily the case that Q if P, then it's necessarily the case that Q if it's necessarily the case that P.

7.5.4 What is a logical constant?

Quine's analysis, we noted, is vitiated by his failing to say what conditions an expression must satisfy to be a "logical constant." We are in position to provide this information. The expressions "not" and "if . . . then . . ." are on Quine's list of logical constants, and those expressions occur in:

(5) "If Smith is at home, then it's not the case that Smith is not at home."

Before we proceed, we have to make a couple of points about the word "not." That word *what* is known as a "sentence-level operator": it's an expression that forms sentences out of sentences—it forms "snow is not white" out of "snow is white." It isn't often that sentence-level operators are replaced with variables. But not only *can* such replacements be made; they must be made. Otherwise, it would be impossible to express any generalization about sentence-level connectives. (Logicians and linguists often make statements about whole classes of sentence-level connectives, and such statements involve replacements of this kind, since they have the form: "given *any* sentence level connective X . . .") In fact, we're about to state just such a principle right now.

Given (5), if we replace "not" with a variable, the result is:

(5X) "If Smith is at home, then it's X the case that Smith is X at home."

If we bind the free variable with a universal quantifier of the appropriate kind, the result is:

(5XUG) "Given any sentence-level operator X, if Smith is at home, then it's X the case that Smith is X at home,

which is false, the reason being that some sentence-level operators lack the relevant property. An example of such an operator is the word "necessarily." When this word replaces the variable in (5X), the result is:

(5XN) "If Smith is at home, then it's necessarily the case that Smith is

necessarily at home,"

which is false.

Another expression on Quine's list that occurs in (5) is "if . . . then . . ." This expression, like "not," is a sentence-level operator. But whereas "not" is a *one*-placed sentence-connective, "if . . . then . . ." is a *two-place* sentence connective. So, whereas "not" assigns sentences to single sentences, "if . . . then . . ." assigns sentences to ordered pairs of sentences. Given the ordered pair <Toby is a dog, Toby is a mammal>, "if . . . then . . ." yields the sentence: "if Toby is a dog, then Toby is a mammal."

Everything just said about "not" is true of it. If it's replaced with a variable, which is then bound by a universal quantifier of the appropriate kind, the result is:

(5IT) For any two-place sentence-connective X <Smith is at home, it's not the case that Smith is not at home>

But (5IT) is simply false. Let C be a two-place sentence-level operator defined as follows: Given two propositions (or sentences) P and Q, C(P,Q) is true exactly if P is incompatible with Q. In that case:

(5C) C<Smith is at home, it's not the case that Smith is not at home>

is false.

What we just said about "not" and "if . . . then . . ." is true of all the other items on Quine's list of "logical constants." And it's true of many other expressions that Quine did not personally include on that list but that, according to many, nonetheless belong on it. (For example, some hold that "necessarily" and "possibly" belong on that list.) This suggests that, when Quine judged an expression to be a "logical constant," he did so on the (cogent) grounds that replacing occurrences of it with variables systematically failed to yield universally true open-sentences.

This suggests that the following definition of "logical constant" might be the right one (we'll soon see that it's only approximately right):

(LC[357]) an expression E to be a "logical constant" is for there to be some universally-true open sentence that contains E.

LC is close to the truth, but it falls just short. Consider the open-sentence:

(#)‹If P and Q and grass is green, then grass is green.›[358]

(#) is universally true. But, since "grass is green" isn't a logical constant (*) is false. But, apart from "P," "Q," and "R," all the expressions occurring in:

(##)‹If P and Q and R, then R.›

are logical constants. (##) is universally true, and it contains no expressions that, when replaced with free-variables, result in some *other* open sentence that is universally true. If a universally true open sentence contains no such variables, let us say that it is "pure." This shows that a "logical constant" is any expression other than a variable that occurs in a pure, universally true open-sentence.

7.6 analytic truth ≠ formal truth

As we discussed in Chapter 1, the two main contentions of Wittgenstein's *Tractatus Logico-philosophicus* (TLP) are (i) that all non-empirical truths are tautologous (definitional truths; e.g., "a fortnight is a period of two weeks") and (ii) that all non-tautologous truths are observation-reports.

In Chapter 1, we saw why each contention is false. Given the points just made, we are in a position to put forth another reason to reject (i) and, therewith, to reject a third contention of the TLP, viz. that all formal truth is analytic truth and, therefore, that all entailment is formal entailment. Both claims are false.

> ‹If P and Q and R, then R› is true for all values of its variables. This means that:

(A) For any propositions P, Q, and R, if P and Q and R, then R.

But, (A) is not itself a formal truth. (A) is true, of course. But it is not in virtue of its being an instance of a universally true open-sentence that it's true. For it isn't an instance of such a sentence. Any open-sentence of which (A) is an instance is one of which:

(B) For no propositions P, Q, and R, if P and Q and R, then R,

which is obviously false. Since (B) is an instance of any open-sentence of

which (+) is an instance, the latter is not formally true.

It is in virtue not only of its form, but also of what is meant by "for any propositions P, Q, and Q" that (A) is true. (A) and (B) have the same form. What distinguishes them is that, in the place where (B) contains an occurrence of one expression ("for *no* propositions P, Q, and R"), (A) contains an expression with a different meaning ("for *any* propositions P, Q, and R").

So the fact that A is, whereas B is not, analytic cannot possibly be chalked up to the fact that one is, whereas the other is not, formally true. Thus, *not all analytic truth is formal truth.*

In the TLP, Wittgenstein[359] says that all analytic truth *is* formal truth. Wittgenstein is wrong, for the reasons just given. Also, even if there were some formally true sentence S such that S was equivalent with (A), the statement:

(C) ‹S is equivalent with A.›

would not *itself* be a formal truth. To be a formal truth is to be a consequence of an informally true universal generalization. The statement that:

(%) If grass is green and snow is white and roses are red, then roses are red

is a formal truth because it is a consequence of (A). Thus, Wittgenstein's thesis that all analytic truth is formal truth is not only wrong but incoherent.

Here is another illustration of that fact. The sentence:

(JTB) any case of knowledge is a case of justified true belief

is an analytic, not an empirical, truth. But it isn't a formal truth, since it has the same form as

(NTB) any case of knowledge is a case of unjustified false belief,

which is false.

Wittgenstein grants that there *seem* to be analytically true sentences that aren't formally true. But he denies that there are really are such sentences. Here is what he says:

(AFT[360]) Though it *seems* not to be formally

true, the sentence

(JTB) “any case of knowledge is a case of justified true belief”

is in fact formally true, the reason being that it is an abbreviation of

(JTBF) “any case of non-accidentally true, justified belief is a case of justified true belief,”

or, at any rate, of some other formally true sentence.

AFT is false. JTB is not JTBF in disguise. They express different propositions. JTB is informative. JTBF is not. One can accept the proposition meant by the one without accepting the proposition meant by the other. JTB and JTBF express *equivalent* propositions (assuming, as we are, that JTBF is correct). But if JTB is to be a mere abbreviation of JTBF, both must express the very same proposition. They don’t; so AFT is wrong.

According to AFT, any given informal analytic truth can be turned into a formal truth by replacing expressions with analytic definitions of themselves —by replacing “circle” with “closed planar figure of uniform curvature,” etc. But such a formalization would *presuppose* the correctness of all the analyses involved. And none of those analyses would be expressed by formal truths. JTBF is indeed formal. But it doesn’t express an analysis. What *does* express an analysis (or so we will suppose for argument’s sake) is:

(JTBA) “any case of knowledge is a case of non-accidentally true, justified belief.”

But JTBA isn’t *formally* true, the reason being there are formally identical false statements (for example, “any case of knowledge is a case of accidentally true, unjustified belief”).

7.7 Conceptual analyses never expressed by formal truths

A ***conceptual analysis*** is a non-empirical, non-trivial sentence of the form ‹C_1 is fulfilled if and only if C_2›, where C_1 and C_2 are conditions. (The

sentence “condition C is fulfilled” is synonymous with “property P isn’t instantiated.”) Thus:

(CK) x is a circle if and only if x is a closed planar figure of uniform curvature

is a conceptual analysis, and so is

(CK*) x is a closed planar figure of uniform curvature iff , for some plane P and some point D on P, x is the area bounded by every point on P that is a fixed distance from D.

Neither CK nor CK* is a formal truth.

No conceptual analysis is formally true. Conceptual analyses are expressed by analytic statements, and analytic statements either: (i) coincide with informal, analytic universal generalizations or (ii) are instances of such generalizations. If (i), they’re obviously not formally true. But any statement that is an instance of some informal, analytic universal generalization is trivial, and is therefore not analytic. For example, % is a consequence of (A), and % is clearly trivial.

PART V

Ethics and Religion

Chapter 19

Some Fundamental Principles Relating to Ethics

1.0 What is ethics?

There are two kinds of statements: ***normative*** and ***descriptive.*** A descriptive statement is one to the effect that such and such is the case. Descriptive statements don’t express value-judgments. Examples of such statements are

"Smith is over 6-feet tall" and "grass is green."

A normative statement is one that *does* express a value-judgment; such a statement is one to the effect that such and such falls either short of, or satisfies, or exceeds some *standard* or *norm.* Examples are "it was evil of Hitler to commit genocide," "it is wrong to steal," "it is commendable to give money to charity," "Smith was acting valiantly in saving the drowning toddler."

A normative category is one that occurs only in normative judgments. Examples of such categories are *good, bad, just, unjust, valiant, noble, wicked, depraved, commendable,* and *condemnable.* Ethics is the discipline that attempts to clarify the structures of these concepts—that attempts to state as clearly as possible the conditions that things must satisfy to fall under them. So ethics attempts to make explicit exactly what conditions an act must satisfy to be praiseworthy or that an institution must satisfy to be just.

2.0 Two kinds of goodness: instrumental and intrinsic

The most important normative terms are "good" and "bad." But here we must be careful. For each of these terms is *ambiguous*—that is, each has more than one meaning, and each has *both* normative *and* non-normative (descriptive) meanings. Let us start with the word "good."

There are two kinds of goodness: *instrumental* good and *intrinsic* good. To say that an act is instrumentally good is to say that it has consequences that are desired by, or to the advantage of, the agent. (The agent is the person who commits the act.) So if my sole objective is to make money, it is instrumentally good, at least as far as my interests are concerned, to steal a million dollars from a local orphanage (provided that I don't get caught or otherwise get into trouble for doing so). So stealing money from the orphanage is *instrumentally* good. That is to say, it is *useful:* it has consequences that are to my advantage; it is in my *practical* interest to act in that way.

At the same time, of course, stealing money from a charity is not *ethically* good. It is good *only* in a practical, instrumental sense. It is not good

in a *non*-instrumental sense.

If something is good in a non-instrumental sense, it is *intrinsically* good. Ethics is concerned with *intrinsic* goodness. It attempts to answer the question "which things are *intrinsically* good?"[361] Ethics is not interested in instrumental (i.e., strategic or pragmatic) goodness. Ethics is not interested in identifying ways that thieves can steal more or ways that crooked politicians can acquire more power. Ethics is interested in identifying those courses of action, and those states of affairs, that are good *even if their consequences are disregarded.*

There are many views as to which things have intrinsic value. But here are some things that many ethicists, though not all, believe to fall into that category: happiness, intelligence, benevolence, honesty (both with oneself and with others), loyalty, pleasure, life (being alive is a good thing, even though life has its less-than-stellar moments), friendship, compassion, enthusiasm for one's work and for life in general, and freedom. It should be pointed out that there are different kinds of freedom: there is freedom from *external,* physical coercion (this is the kind of freedom one has in virtue of not being in a jail cell); there is emotional freedom (this is the kind of freedom one has in virtue of not being dominated by internal, psychological compulsion); and there is also intellectual freedom (this is the kind of freedom one has by virtue of being able

to figure things out for oneself, as opposed to slavishly relying on hearsay). I leave it open whether there are other kinds of freedom.

Helping a person in need, in cases where doing so is not to one's own practical advantage, is intrinsically good. Having an expansive consciousness is intrinsically good. Playing a musical instrument is intrinsically good. (It is also instrumentally good—no pun intended—since, as psychologists have recently discovered, it promotes neural plasticity.) Being honest with people, when it would be more convenient to lie, is intrinsically good.

I wish to emphasize that it would be hard to find two ethicists who were in complete agreement as to which things were intrinsically good. But the tentative, and perhaps partial, list just given makes it clear what, at least approximately, is meant by the expression "intrinsic good."

2.1 Instrumental and intrinsic goodness not mutually exclusive

Having a sharp intellect is both an intrinsic good *and* an instrumental good. Having intelligence is an instrumental good, since it can obviously improve one's strategic position and, more generally, can have good consequences. But the goodness of being intelligent exceeds any practical advantage that it might bring to one. What we just said about intelligence is true of happiness, of pleasure, of friendship, and of honesty. All of these things can be instrumental goods, since they can have good consequences. (Of course, they can also have bad consequences: sometimes the worst thing you can do is be honest. But, undeniably, honest is *sometimes* an instrumental good: in general, if people know you're truthful, they're more likely to help you out.) At the same time, all of these things are good in and of themselves. To take another example: being happy tends to give one the confidence to act in ways that are to one's strategic advantage. But the value of happiness obviously exceeds the strategic advantages it may bring one. Bottom line: something can be both instrumentally and intrinsically good.

3.0 More meanings of the word "good': goodness vs. commendableness, badness vs. condemnableness

We just saw that the expression "good" is ambiguous between "instrumental good" and "intrinsic good." But the expression "intrinsic good" is *itself* ambiguous. (It is common in philosophy to encounter ambiguities within ambiguities.) Some things that are intrinsically good are *commendable*—that is, they deserve *praise.* And some things that are intrinsically good *do not* deserve praise. In, fact, as we'll soon see, some intrinsic goods, paradoxical though it may sound, deserve *censure*—that is they are to be *disapproved of.*

For example, happiness is intrinsically good, as we just noted. Of course, being happy may have good consequences. But a person's happiness has a goodness that cannot be identified with the good consequences, if any there be, that it brings to that person.

But does a happy person necessarily deserve *praise* for being happy?

Supposing that Smith is somebody who is just naturally chipper, it's obviously a good thing—nay, an intrinsically good thing—that Smith is in good spirits. The more happiness, the better, after all. But Smith doesn't necessarily deserve any *praise* for being happy. It could be he is just naturally happy and that he doesn't have to make any special effort or do anything particularly estimable, to remain happy or, if he happens to become blue, to become happy again. Of course, it's *possible* that Smith has to commit noble deeds to be happy—that he's deeply depressed *unless* he has recently done something good. It's possible, for example, that Smith will fall into a deep depression unless he has just saved the life of a drowning child or helped a child learn to read, or some such. But it's also possible—and, if Smith is like most people, fairly probable—that his being happy is not in general a consequence of his doing something praiseworthy. Maybe Smith's endocrinal system just happens to manufacture happy hormones all the time; maybe Smith is happy because he's very shrewd at sidestepping the various social and professional landmines that less cunning people are likely to step on; maybe Smith is evil and he has found a way to commit evil acts, and thus to make himself happy, without getting in trouble. There are a lot of reasons why

Smith might be happy. And, undeniably, Smith's happiness is a good thing. But it is by no means something that is *in all cases* praiseworthy.

To take a more stark example, the happiness of a rabbit is a good thing. But, given their cognitive limitations, rabbits are not capable of acting in a praiseworthy manner. To be sure, it's *good* that they're happy (when they're happy). But, good though a rabbit's happiness is, the rabbit doesn't deserve praise for being happy—it isn't in the same category as Gandhi or Mother Teresa. So happiness is non-instrumentally good, while at the same time not being *commendable.* ("Commendable" is a synonym for "praiseworthy.")

But, of course, some intrinsically good things *are* praiseworthy. Suppose that Jones spends a great deal of time and energy working for a charity and that he does so knowing he does not reap any kind of practical benefit from his labors. Jones' behavior is obviously praiseworthy—that is, it is *commendable,* and it is also intrinsically good.

Another example: Suppose that Smith is a naturally talented musician

who plays the piano very well. Suppose, in fact, that Smith is blessed with so much native talent that he never had to work hard to play the piano well—it just came to him naturally. Smith's musical talent is intrinsically good, as is the happiness it brings him. But Smith doesn't deserve any *praise* for happening to have been blessed with this talent. (How could he deserve praise for the occurrence of something over which he has no control?) It's good that he has such talent. But he doesn't deserve to be praised for it in the way in which Mother Teresa deserves to be praised for her good works.

Bottom line: there are two kinds of intrinsic goods—those which are commendable and those which are not.

3.1 Two kinds of intrinsic badness

Just as there are two kinds of intrinsic goodness, so there are two kinds of intrinsic badness.

It's obviously bad to be unhappy. And the badness in question is clearly of a non-instrumental, and therefore intrinsic, nature. In any case, even if unhappiness is instrumentally bad (i.e., even if it has bad *consequences* for those who suffer from it) it is also, quite clearly, bad in and of itself: being unhappy is bad even if it doesn't bring about misfortunes *other* than happiness. So even in cases where unhappiness has bad consequences (e.g., it might weaken one's immune system or alienate potential benefactors), it is *also,* in such cases, intrinsically bad.

At the same time, one does not, at least not in all cases, deserve to be *censured* for being unhappy. Somebody who is grieving the loss of a loved one doesn't deserve to be reprimanded for having that emotion. If anything, if somebody *failed* to grieve the loss of a loved one, they ought to be censured for *that.* (Only a monster would feel nothing in response to the death of a beloved child or spouse.)

That said, some intrinsically bad things obviously *do* warrant censure (i.e., they are *condemnable*). Causing somebody else to suffer, for no good reason, is intrinsically bad *and* it is condemnable.

Bottom line: some, but not all, intrinsically bad things warrant censure (i.e., are condemnable); and, therefore, "intrinsically bad" is not, at least not always, identical with "worthy of condemnation."

4.0 The morally complex structure of some situations

A single act or state of affairs can be intrinsically good in one respect and intrinsically bad in some other respect. Suppose that I am playing the piano, that I'm doing so very well, and that I'm taking pleasure in it. Surely that is an intrinsic good. But suppose that, because I'm playing the piano, I am waking up my very sick roommate, who needs his sleep, and that I am aware of the harm that I am doing to my roommate. In that case, my piano playing, though an intrinsic good, is condemnable. It is wrong, but it is also intrinsically good. This may seem like a paradox, but it isn't. The reason it isn't a paradox is that, as we'll see in a moment, "wrong" is not the opposite of "good"; in other words, wrongness is not the absence of goodness. And this is closely related to another fact that we'll soon discuss, namely, that "right" is not identical with "good."

A given act can be *both* condemnable *and* commendable. Suppose that Jones, who is an exceptionally tough and vindictive fellow, insults Smith and that Green, who is a friend of Smith's, savagely beats Jones to avenge Smith's honor. (Let us assume that Green's act is motivated only by loyalty to Smith and there is no element of self-interest in it.) Green's act is commendable, because it is an act of loyalty and courage. But it is also condemnable, because it is an act of excessive and unnecessary violence.

5.0 The non-privative character of moral attributes

Despite what first appearances might suggest, there is nothing paradoxical about the fact that a given act can be both commendable and condemnable. This is because for something to have the property of being commendable is not for it to *lack* the property of condemnable, and for something to have the property of being condemnable is not for it to *lack* the property of commendable. Neither property is "privative"; they are both "positive" properties. A "privative" property is one that a thing has when it's *missing* something. A "positive" property is one that a thing has when it has an *extra* something.

For something to be cold is for it to *lack* heat. For something to be dumb is for it to *lack* intelligence. For something to be poor is for it to *lack* money. To be dumb or cold or poor is to *lack* something. There are many properties like this: in other words, where many properties are concerned, having them is identical with *failing* to have some other property.

A property that is *not* privative is "positive." For something to be hot or smart or rich is not for it to *lack* anything. On the contrary, it is for it to *have* something. That is why these are "positive," as opposed to "privative," properties. Another illustration: for a drink to be sugary is for it to *have* something extra in it. That is why the property of being sugary is positive, as opposed to privative. By contrast, for a drink to be sugar-free is for it to *fail* to have something; that is why the property of being sugar-free is privative, as opposed to positive.

The property of being condemnable is not privative; it is a positive property. (It is, of course, not positive in the *moral* sense. It is positive in a strictly logical sense that has no specific connection to morality.) Just as the property of being poisonous is in no way privative and is therefore a positive property (in the relevant, strictly logical sense of the word "positive"), so the property of being condemnable is in no way privative and is therefore a positive property (same qualification). And the property of being commendable is *also* positive. Thus, the presence of neither property is identical with the absence of some other property; more specifically, the presence of neither is identical with the absence of the other. That is why both properties can be present in some one thing (in some one act or state of affairs), which is why a given act can be *both* commendable *and* condemnable.

People tend to describe acts as *either* "good" or "bad," and they have a certain resistance to characterizing them as *both* good *and* bad. This, I think, is partly because people make the mistake of thinking that badness is the *absence* of goodness. It is not: badness, like goodness, is the *presence* of something. A rock *lacks* goodness. But rocks aren't *bad.* They're neither good nor bad. To be bad, it isn't enough to *lack* something; one must have something extra. Hitler was bad. And he was bad because he had something that rocks don't have—he had ill-will and a willingness to act on it. So it wasn't Hitler's *lacking* something that made him bad. On the contrary, it was

his having an extra something that did so. To sum up, people tend to think that good and bad cannot co-exist in the same person, or the same act, because they wrongly think that good is the absence of bad and that badness is the absence of good.

But there is another, probably more important reason why people fail to appreciate the highly ambiguous nature of morality. People tend, especially where moral and emotional issues are concerned, to think in an unrealistically binary way. If I had to hazard a guess as to why this is so, it would be that it's simply easier—less demanding of energy, intelligence, and reflection—to describe other people either as monsters or as heroes, and that it is therefore easier to take very simple and uniform positions with respect to their actions. (Also, by thinking in binary terms, one shields oneself from ugly and painful realities about oneself and about others. It's emotionally easier to see the world as divided into saints and ghouls than it is to see everyone, including oneself, as part-ghoul, part-saint.) In any case, what is important here is that this desire for simplicity, whatever its psychological basis, is not consistent with the actual moral structure of most acts.

6.0 A corollary of the non-binary nature of moral attributes: Moral obligations have a dimension of weight

Moral obligations are not "binary"; they come in degrees. And some obligations are therefore stronger than others. If I promise to meet you for lunch, I have an obligation to show up. But that obligation is not as strong as my obligation to show up to the classes that I'm paid by this university to teach. And the latter obligation is not as strong as the President's obligation to serve the interests of his 300 million constituents.

A related point is that there are degrees of wrongness. It is wrong of me not to show up to a lunch date I promised to show up to; it is more wrong of me not to show up to a class I'm supposed to teach; and it is even more wrong for a head of state to violate the trust of his constituents.

6.1 For one moral obligation to *outweigh* another is not for it to *cancel* it

Oftentimes, we have conflicting obligations. The following story will make this clear. My best friend, who has always been kind to me, has one million dollars in a safe to which I have access. Of course, I have a moral obligation not to steal that money; for, were I to steal it, I'd be harming somebody who's been very decent to me, and also violating that person's trust in me. Unfortunately, I may not have the option of honoring this moral obligation. I have five young children, all of whom desperately need to undergo very expensive medical treatment. By the time I could legitimately make the money needed to fund their treatment, they would have died, and the only way to get my hands on the needed funds is to steal my best friend's money. So I have a choice: steal the money or let me children die. There is no third option.

I obviously have a moral obligation—one that is *even stronger* than my obligation not to steal my friend's money—to get them the care that they need. I'm in a real jam here. If I steal my friend's money, I'm violating his trust and friendship. If I *don't* steal his money, I'm violating my responsibility to help my children.

Under the circumstances, the right thing to do—that is, the *least bad* thing, the lesser of two evils—is to steal the money. After all, the moral value of saving five children exceeds the moral value of not stealing my friend's money. But that doesn't mean that my obligation not to steal my friend's money ceases to exist. It *does* still exist. In stealing from my friend, I am doing something to him that he has a right *not* to undergo and that I therefore have an obligation not to make happen. Basically, obligations that are outweighed by other, more powerful obligations do not on that account cease to exist.

6.2 "right" ≠ "good", "right" = "least bad"

Under the circumstances described a moment ago, the "right" thing to do is to steal my friend's money, the reason being that, if I don't do this, my children will die. But stealing my friend's money is clearly *the lesser of two evils.* In other words, it is a bad thing, a terrible thing—but the alternative is even worse. So what makes it right to steal my friend's money is not that doing so is a positive good; it is that doing so is the *least bad* option available under the circumstances. In general, what we mean by "the right course of action"

is the *least bad* course of action that the circumstances permit.

6.3 "wrong" ≠ "bad", "wrong" = "least good"

For similar reasons, the "wrong" thing to do is not necessarily a *bad* thing to do. For a course of action to be wrong is for it to be the *least good* possible course of action. Suppose that I am in a situation where, for some reason or other, there are only three possible courses of action available to me: (i) I can save the lives of individuals M_1 . . . M_{10}; (ii) I can save the lives of individuals M_{11} . . . M_{15}; (iii) I can save the lives of individuals M16–M17. In other words, my choices are (i) save 10 people, but let 7 people die; (ii) save 5 people, but let 12 people die; or save 1 person, but let 16 people die. (Let us suppose that the individuals composing any one of

these groups are no better or worse, no more and no less worthy of life, than the individuals composing either of the other two groups.) If I save the lives of M_{16}–M_{17}, I am obviously doing a good thing; for it is obviously good to save lives. But I'm still doing the *wrong* thing, since it would obviously be better to save M_1 . . . M_{10}. After all, the more people saved, the better (assuming, as we are, that the relevant individuals are all of comparable value).

7.0 "Ought" implies "can"

It is a generally accepted principle of ethics that one has an obligation to do something only if one *can* do it. One ought to do something only if one can do it—"ought" implies "can." Here's the idea. Because it isn't within my power to heal every sick person on the planet, I have no obligation do so; and I have an obligation to treat my students fairly only because it *is* within my power to do so.

There is much truth in the principle that "ought" implies "can." But, in light of what we said a moment ago, this principle must be taken with a grain of salt. We saw that there are circumstances where one has multiple conflicting obligations—where one has an obligation to X *and* Y *and* Z, but where it isn't possible for one to do *all* of those things. Under such

circumstances, one *ought* to do something (namely, X *and* Y *and* Z) that one cannot do, showing it isn't strictly correct to say that "ought" implies "can."

But that principle has not been proven completely false. For, under the circumstances just described, there is no *single* course of action that one ought to carry out that it isn't within one's power to carry out. It *is* in my power not to steal my friend's money (even though it's not in my power to discharge *that* obligation *and* my obligation to save my children); and it does seem that, if that conditions weren't met, I wouldn't have an obligation not to steal my friend's money.

8.0 Legality ≠ morality

There are, or at least can be, acts which are legal, even legally required, which are immoral. In this country, before the Civil War, it was legal to have slaves, and it was *illegal* to set another person's slaves free. But it is exceedingly immoral to have slaves and exceedingly moral to set slaves free.

At the same time, legality and morality often overlap. It is illegal to kill people (except in self-defense), and it's also immoral to do so (same qualification). It is a delicate question how exactly legality and morality are related. See Chapter 27 for a discussion of this issue.

9.0 The moral status of passing judgment on X ≠ to the moral status of X itself

Given only that some object *in fact* weighs 200 pounds, I don't necessarily know, or therefore have the right to say, that it has that weight. After all, I might not have enough information to make that judgment. Given only that some statement is correct, I don't necessarily have the right to make it. I have that right only if I have good reason to believe that it is correct. And I may simply not have enough evidence to make that call.

This principle has an important ethical consequence. Jones, let us suppose, is on trial for a murder that he *in fact* committed. But the jury doesn't have enough information to convict Jones. The jury just doesn't know whether Jones is guilty. Jones' act is condemnable. Murder is wrong, after all. But the jury doesn't *know* that Jones committed murder; and the jury therefore doesn't have the right to convict Jones. It would be condemnable of

the jury to do so. It is wrong to convict a person that one doesn't know to be guilty, even if that person is guilty.

For similar reasons, even if an act *is* condemnable, one doesn't necessarily have the right to describe it as such. One has that right only if one has sufficient information about the act. If, without having such information, one condemns it, then one is oneself acting condemnably; for one is making an accusation that one doesn't know to be correct, even though, by coincidence, it *is* correct. Bottom line: it can be condemnable to condemn the condemnable.

10.0 The goodness of the act vs. the goodness of the agent

Whether somebody deserves to be commended for an act depends, not on whether it actually was the right act, but on whether it was the agent's intention to perform the right act. And whether somebody deserves to be reprimanded for an act depends, not on whether it actually was the wrong act, but on whether it was the agent's intention to perform the wrong act.

Suppose that Kathy is a very decent person who is a mother of five. Because she's read all the right parenting magazines, she knows that feeding her children nutritional supplement X will make them healthy, happy, etc. So, even though X is expensive, she works extra hours, buys it, and feeds it to her children. And, as a result, they are happy, healthy, etc.

Obviously Kathy deserves commendation. She did the right thing; moreover, it was her intention to do so—that is, it didn't happen by accident.

Now let's consider a different case. Like Kathy, Mary is a mother of five who wants to benefit her children and reads all the right parenting magazines. Given the evidence at Mary's disposal, the most reasonable conclusion for her to draw is that she should feed nutritional supplement Y to her children. It's expensive; but Mary works extra hours to pay for it. And she buys it and feeds it to her kids. Bottom line: Mary is psychologically just like Kathy. Her intentions are exactly the same as Kathy's. So far, so good.

But there's a problem. It turns out that, despite all the advertising and the FDA-run tests, Y is lethal. (It doesn't kill instantly, taking years to do so, and that is why it wasn't previously realized how dangerous it was.) So, after a

couple of years, all of Mary's children die.

It's obvious that Mary didn't do "the right thing." In other words, she didn't act in a way that maximized human welfare and minimized violations of human rights. But is Mary *morally* worse than Kathy? In other words, is Mary to be *judged* more harshly than Kathy? In fact, is Mary any less worthy of praise than Kathy?

No. Mary's intentions are exactly like Kathy's. The difference between the two is that Mary, through no fault of her own, was fed faulty information. Mary, like Kathy, did everything she could to bring about a morally optimal outcome. The difference between Mary and Kathy is that, while the world met Kathy half-way, it didn't meet Mary half-way. For that reason, they are equally worthy of praise, even though the one, but not the other, made something happen that should not have happened.

Third case scenario: Betty is a mother of five. Her intention is to kill her children. (She finds that being a mother is boring and wants to be done with it.) On the basis of careful research, she concludes that feeding her children substance Z will kill them and, moreover, won't be detectable by forensic scientists. (Z is a toxic powder that looks and tastes like sugar.) So she feeds Z to her children. But what the researchers didn't know, and what Betty therefore never found out, is that, although Z is usually fatal, it has extremely positive health effects when consumed by children of a certain very rare blood type. Betty's children happen to be of this blood type. So, because they consume Z, they live happily ever after.

Betty brought about a good state of affairs—one where everybody (except herself) was happy. But does Betty deserve praise? No. In fact, Betty clearly deserves condemnation of the most severe kind. But how can that be, given that she brought about a good state of affairs? Because, even though she ended up doing good, it was not her *intention* to do so. It was her *intention* to do evil. So she deserves no praise and in fact deserves condemnation.

Thus, whether an agent deserves praise or censure depends, not on what she or he *in fact* brings about, but on what it is her/his *intention* to bring about.

11.0 Why many legal systems (including ours)

punish attempted crimes less severely than successfully carried out crimes

Attempted murder is punished less severely than actual murder. Why? Because failure to carry a plan out successfully oftentimes (though not always) suggests that the person's heart wasn't really in it. If Smith is a smart guy who flunks out of law school 10 times, it's a distinct possibility that Smith's heart isn't really in it. If somebody tries to commit murder but fails, it's a possibility that he was of two minds about it. And that might explain why Smith failed to carry out the crime. When he pulled the trigger, let us suppose, his gun jammed. But his gun wouldn't have jammed if, before trying to commit the crime, he had made sure that it worked. Had it been Smith's resolute intention to carry out the crime, then he might, very possibly, have made sure that his gun did work. And, supposing that Smith did in fact have reservations about committing the murder, it's a distinct possibility that those reservations were of a moral nature. (It's also possible, of course, that those reservations were of a purely strategic nature.)

So failure to carry out a crime sometimes, though obviously not always, results from a certain unwillingness to go through with it; and such an unwillingness sometimes, though not always, results from *moral* reservations that one has about the crime.

The reason, therefore, why the criminal justice system does not punish would-be murderers as severely as actual murderers is that would-be murderers oftentimes are not as intent on committing murder as actual murderers. So the criminal justice system's position here is consistent with the view, argued for previously, that it is agent's intention, as opposed to the result of that intention, that determines how he ought to be judged.

For similar reasons, failure to go through with some projected good deed may show that one does not really want to perform it. Let us suppose that you agree to visit your ailing aunt, who loves your company but who you find extremely tiresome. Every time you are scheduled to visit her, there is some kind of problem. You lose your car keys. You are overwhelmed by a desire to clean your house. You finally find the inspiration to write the first chapter of your novel. These problems are real: you really do lose your car keys; you really do, at long last, know how to begin your novel; and so on. And you are

not responsible, at least not consciously, for any of these last-minute problems. They just seem to spring up. Nonetheless, the fact that they always do spring up, taken together with the fact that you really hate visiting your aunt, is suggestive. It suggests that, deep down, you *don't* want to visit your aunt. It suggests that your intention to visit your aunt is a feeble one.

This explains why people who *succeed* in carrying out projected good deeds are given more commendation than those who simply attempt to carry them out. Failure to carry out a good deed often (though not always) results from a lack of willingness to carry it out—from, in other words, an insufficiently resolute good intention. At some level, people are aware of this fact, that being why would-be do-gooders receive less commendation than actual do-gooders.

So while it's true, as we saw in the previous section, that two people who have *precisely the same intentions* deserve precisely the same amount of praise or blame, even if one of them succeeds in carrying out his intentions and the other fails, it's still probably a good idea, as a general rule, to punish attempted murderers, arsonists, etc., less severely than their successful counterparts. For, in general, those who *successfully* carry out these crimes are less conflicted about whether they want to succeed in their nefarious endeavors.

12.0 Aren't there people who are to be condemned *even though* it is their intention to do good?

It is generally agreed that Hitler's acts were wrong and that Hitler deserves to be condemned for them. "But," it is often asked, "wasn't it nonetheless Hitler's intention to do good? His efforts were, of course, misguided. But that doesn't mean that he didn't have good intentions. And, supposing that he did have good intentions, isn't that inconsistent with your [Prof. Kuczynski's] point that it is on the basis of one's *intentions,* not one's deeds, that one deserves condemnation?"

My response is that it pretty clearly was *not* Hitler's intention to do good. (In any case, if he did have such an intention, it was grossly outweighed by other intentions.) Obviously when he was giving speeches, he didn't say "it's my intention to do evil. Thanks a lot, you've been a great crowd." But that

tells us nothing about what his intentions really were. And given how obvious it is to anyone of normal, or even minimal, intelligence that gratuitous torture and murder are wrong, it's not feasible to suppose that, as a rule, Hitler's intentions were good ones.

13.0 Can one act immorally towards oneself?

This is a controversial issue; but my feeling is "yes." One can sell oneself short; one act self-destructively. One can do things to oneself that it is one's right not to undergo.

There are many ways to do this: being too easy on oneself (permitting oneself every kind of excess imaginable: drugs, binge-eating, etc.); being too *hard* on oneself (pushing oneself *too* hard, setting excessively high standards). This list is far from complete.

This is not to say that people who hurt themselves are ethically in the same category as people who hurt others. People who hurt themselves are often (though by no means always) hyper-principled, whereas people who hurt others are often (though by no means always) lacking in principle.

"But you just said that people who hurt themselves are behaving immorally. And that is surely inconsistent with your statement that such people are, at least sometimes, hyper-principled." Given only that an act is principle-driven, it doesn't follow that it isn't wrong or even evil. There are many kinds of ethical principles—and not all of them are good. For example, some people felt that it was their ethical duty to support Hitler, no matter what. This was *an* ethical principle—it just wasn't a good one. By acting in accordance with similarly misguided ethical principles, people may hurt themselves.

It seems true, as a general rule, that people who hurt others deserve to be punished, whereas those who hurt themselves do not. It is concluded from this that self-destructive behavior is not *immoral,* even though it may be bad in some other way.

This argument is less than probative. Whether somebody deserves punishment is not a function *solely* of whether they've done wrong. If the good done by punishing someone is grossly outweighed by the bad that it does, then it may, at least in some cases, be ethically wrong to punish them.

Suppose that little Timmy stole a cookie. It was not good of him to do this. But suppose that, because of difficult life circumstances, little Timmy is emotionally very fragile and needs TLC (tender loving care), and that, if punished, he would completely break down. Under this circumstance, it would probably be better to give him a pass.

If we were to *punish* people who are *already* hurting themselves—who are already punishing themselves—we'd be adding insult to injury. First of all, self-destructive behavior is its own punishment; so an externally-imposed punishment would be redundant. Second, by punishing such a person, we'd be violating their autonomy (their right to determine their own fate), which might be ethically worse than they did.

In Chapter 22, it is argued that *after a fashion* all acts of immorality are acts of immorality towards oneself.

14.0 Some metaethical principles

Metaethics is the discipline that says what, if anything, ethical statements mean. A "valuative" statement is one that expresses a value-judgment; that is, it is one that says something should (or should not) be done, or that something is (or is not) good. Basically "valuative" means the same thing as "ethical" or "having ethical content."

We'll now discuss six metaethical principles. Principles 4–6 are correct; principle 3 is approximately, though perhaps not exactly, correct; and principles 1 and 2 are false, even though they are very popular to this day.

14.1 You can't derive an "ought" from an "is"

This means that, if you had a list of all the true non-valuative statements, you could not, on that basis, justifiably make any valuative statements. You could not say "it is obligatory that you do (or not do) x."

Though widely accepted, this principle may well be false. According to one widely accepted moral doctrine, "x is a good act" means that x makes people happier than the alternative possible courses available. According to another moral doctrine, it means that x promotes the actualization of human potential to a greater degree than the alternative possible courses of action available."

If either of these doctrines, or anything at all like either of them, is on target, then an "ought" *can* be inferred from "is," since "x is what ought to be done" *can* be inferred from "x is more likely than the other possible courses of action to maximize happiness/the actualization of human potential."

14.2 Moore's "open question" argument

Moore argued that, no matter how many truths of a purely descriptive, non-moral nature you know, you don't have enough information to make any moral judgments.

Here is his argument. Premise: No matter how many *valuative* truths you know about x, it remains an "open question" whether x is good or not. Conclusion: Ethical truths are not identical with, or otherwise "reducible to," factual statements.

Analysis of Moore's argument: Premise 1, as we just saw, is false. (In any case, it is false if there is any truth to the thesis that acts are good to the extent that they maximize happiness or enable people to develop their inborn potential.)

In any case, Premise 1 is deeply implausible. If I know that x is a person who kills others for fun and delights in their pain and never does anything for anybody, then it is *not* an open question whether x is good. Bottom line: Moore's argument is no good.

14.2.1 Commentary on the last two principles

If Jones tortures babies for fun, feels no remorse about it, frames somebody else for his misdeeds, etc., Jones' acts, and Jones himself, are obviously bad. So why did Moore think that, even if we have all of this information, it's an "open question" whether Jones' acts are good or bad?

I think it's because, like many other philosophers, Moore didn't have a very good grasp of the concept of entailment. It's often said that P entails Q only if there is nothing in Q that isn't in P. Entailment, it is said, is "non-ampliative." It doesn't add to what you know.

But this is false. Entailment *is* ampliative. To be sure there are *some* cases where it isn't (e.g., "if P and Q, then Q"), but these are more the exception than the rule. The fact that 1 + 1 = 2 entails that there are continuous functions that cannot be differentiated at any point. It also entails

that space is metrically amorphous. This is not to mention that, if we were barred from making deductions, it would be impossible to organize our raw experiences into organized theories. So if it were really true that deduction doesn't increase knowledge, it would follow that people who draw inferences know no more than those who don't and just live in the moment—that our Pleistocene ancestors are as knowledgeable as we are. But that's absurdities. And yet the conception of entailment underlying Moore's argument, as well as Hume's, generates precisely these absurdities. See Appendix 1, Sections 3.1–3.9 for a discussion of why, contrary to what is often said in textbooks on critical thinking, entailment/deductive inference is ampliative. Also see Chapter 1, Sections 6.0–7.0, and Chapter 18, Sections 7.0–7.7.

14.3 "Ought" implies "can"

In other words, if you can't do it, you have no obligation to do it. If the Earth is about to be hit by a giant meteor, I have no obligation to stop it, since I *can't* stop it. If I could, I probably would have such an obligation. But I cannot reasonably be expected to do what doesn't lie within my power.

Unlike (1) and (2), (3) is at least partly correct—maybe even 100% correct. But there is still *some* reason to doubt it. Suppose that you give me $10,000 to move some rocks for you. Before I move the rocks, I spend the money. Then, when I try to move the rocks, it turns out I can't. I'm not strong enough (or my truck doesn't have enough horse power, or whatever). It still seems that I have an obligation to move the rocks (or to refund the money—but maybe I can no more refund the money than I can move the rocks). In any case, (3) is at least approximately correct.

14.4 Genetic questions must be distinguished from normative questions

Suppose that, as legend has it, Newton came up with the inverse square law (the law that shows how the gravitational attraction between two bodies can be determined on the basis of their masses and the distance between them) as a result of being hit on the head by an apple. The fact that Newton's hypothesis had such a random and ignominious origin is irrelevant to whether

it is correct or not. Whether an idea is correct has to do *solely* with whether it corresponds to the facts. *How* it came about, and who came up with it, are totally irrelevant. Of course, if somebody comes up with an idea as a result of a logical process of reasoning, that may make it more likely than not to be correct. But an idea's being generated in a logical manner isn't what *makes* it correct. What *makes* it correct is its fitting the facts. And ideas that are generated in a logical manner can fail to fit the facts, and ideas that are generated in an illogical manner can fit them perfectly well. Oftentimes, of course, people *do* judge beliefs correct or otherwise meritorious on the basis, not of how well they fit the facts, but on the basis of how they originated. In making such a judgment, one commits the *genetic fallacy.*

How does all this relate to morality? Suppose, for argument's sake, that our ethical views derive from ancient religious beliefs that are no longer held by anyone. Suppose anthropologists discovered that our ethical views on torture (e.g., that torturing infants is wrong) were ultimately derived from a belief that was widely held 500,000 years ago to the effect that the Great Volcano god abhorred infant torture. Would that invalidate our belief that torturing infants is wrong? No. It would show that our *grounds* for having this belief are not good ones. (In any case, it would show this *provided* that our original grounds for having it—viz. our beliefs concerning a certain non-existent Volcano god—hadn't been replaced with other, more logical reasons for believing it.) But people can have very bad grounds for perfectly correct beliefs, and they can have very good grounds for wrong beliefs. So the question "is belief X true?" must at all costs be separated from the question "what is the origin of belief X?"

There is another point of relevance. (This point develops a parenthetical remark made a moment ago.) Oftentimes one's original reasons for believing something are irrational or, at any rate, less than cogent. There are many reasons why one might believe that the Earth is round. Suppose that 3-year old Tim believes it because he likes the Earth and, since he also likes round objects, he infers that the Earth must be round. That would be a bad reason for believing that the Earth is round. But the belief itself is still correct. And, in due course, Tim may replace his initial, bad reason for having that belief with a good reason. So, on the basis of reliable reportage and astronomical calculations, Tim comes to have a *rational* basis for having his correct—but

previously illegitimately arrived at—belief concerning the shape of the Earth.

This happens a great deal in morality. Many of our moral beliefs are simply *forced* on us, in the process of our being acculturated. One is punished if one steals, and not punished, and maybe even rewarded, if one doesn't. The only beliefs that have legitimate foundations are those that are based on evidence and reason; and no belief that one is forced to have, or that one is conditioned to have through a system of punishments and rewards, has such a basis. Nonetheless, those beliefs of ours that result from cultural conditioning sometimes turn out to be justifiable and even correct. So *given only* that one is coerced or conditioned or even

brainwashed into having a certain moral position, it doesn't follow that that position is false. (Nor does it follow that it's correct.) And this is because, as we said earlier, *how* a belief is arrived at is separate from whether it's true.

That said, views that are *in fact* arrived at for reasons having nothing to do with God, religion, or tradition are often *claimed* to have divine origins. Whenever somebody says that God told them that He wants such and such, we have no way of authenticating such a claim, and no reason to believe it, *independently* of our preexisting views as to the merits of such and such. So given a case where it is generally believed that God told so and so that x is good (or bad), it's only because people already had views as to x's merits that they're willing to believe that God actually said that to so and so. (If you tell people that God told you to destroy hospitals, nobody will believe you, the reason being that people already believe that destroying hospitals is evil and, therefore, that God wouldn't say otherwise. If you tell people that God told you to be kind to others, a few people might actually believe you, the reason being that people already believe that being kind to others is good and, therefore, that God might actually say as much.) And those views will in many cases, though not all, be of a logical kind. But, if you convince people that God told you that x must be done, you'll get a lot more takers than you will if you tell people that your Herculean intellect told it to you. For this reason many supposed instances of the genetic fallacy aren't instances of it at all.

14.5 Questions concerning the *use* to which

supposed moral truths are put must be separated from questions concerning morality *per se*

There is no denying that many moral statements are pure propaganda—that they are cynical manipulations the purpose of which is, not to promote the interests of justice, but to advance the private agenda of some nefarious party. Sometimes *correct* moral statements can be put to evil use. There have been many occasions when, in order to give credibility to some evil program, a politician invoked a legitimate ethical principle. For example, Stalin, in his efforts to become the Supreme Leader of the Soviet Union, often talked about how important it was to have mercy for one's political rivals. Of course, being merciful *is* a good thing, but it doesn't follow that Stalin was doing a good thing *in citing that principle.* For his citing it was deceptive, even though the vehicle for the deception was a correct (or, at least, a plausible) one.

At the same time, much of what is represented as moral truth is moral falsehood. What people say is right may be wrong, and what they say is wrong may be right. But it doesn't follow that there's no moral truth. Given only that what is *alleged* to be moral truth (e.g., by self-interested actors) isn't moral truth, it doesn't follow that there is no moral truth—just as, given that alleged physical laws aren't always actual physical laws, it doesn't follow that there are no physical laws.

14.6 Questions concerning morality must be separated from questions concerning the implementation of moral principles

It is often argued that morality has no objective basis, the reason being that, supposedly, it's a human creation. Why is it thought to be a human creation? Because, we are told, morality came into existence a few thousand years ago. At some point, hominids started regulating their conduct in accordance with moral principles. And it is inferred from this that it was at that point that morality was created. Basically, the argument is that, since there was a time when people started being moral, it follows that morality is a human creation

and therefore has no objective basis.

This argument involves two fallacies. There was a time when people first became aware of the fact that the Earth is round; but the Earth was round long before then. Similarly, given only that, at some time, people became *aware* of moral truths, it doesn't follow that moral truths were *invented* at that point.

The other mistake this argument makes is to confuse moral laws with their social implementation. Moral principles are not social rules. Laws in the moral sense are not like laws in the legal sense. There can be no laws in the legal sense in the absence of social institutions. But even in a context where there is no society, it is still wrong to gouge somebody's eyes out for fun. The social institutions that *implement* moral principles—that *enforce* them—obviously came into existence at some point in time and also obviously depend for their existence on human artifice. But it doesn't follow that moral principles themselves came into existence at that point. It wasn't until the 1800s that machines (calculators) were built that *implemented* the laws of arithmetic; but those laws existed long before those machines ever did. So even though those machines are the products of human artifice, it doesn't follow that the same is true of the principles they implement.

15.0 Some bad reasons to reject ethical realism

There appears to be less agreement in ethics than in other disciplines. From this it is concluded that ethical statements merely express "opinions" that have no objective basis and, consequently, either don't mean anything or are categorically false.

But there are at least four reasons why this is not good reasoning. (i) There is a great deal of disagreement where strictly factual issues are concerned. (ii) Even though there will be *more* disagreement in ethics than in other disciplines, that doesn't show that ethical statements always lack an objective basis; for at least certain kinds of disagreement actually *presuppose* the existence of objective fact, and there is some evidence that ethical disagreement falls into this category. (iii) There *is* a great deal of agreement where ethical issues are concerned, the belief to the contrary being based largely on misleading appearances. (iv) Much of the disagreement within

ethics concerns matters of fact, as opposed to matters of value.

Let us discuss each of (i)–(iv).

15.1 Discussion of (i)

Consider the statement:

(KV) If Kennedy hadn't been assassinated, the United States would have pulled out of Vietnam before 1970.

KV is not a valuative statement—it doesn't express a value-judgment. It affirms the existence of a causal relation. It is to the effect that, for reasons of natural law, one state of affairs (Kennedy's continuing to be President) would bring about another (the United States' pulling out of Vietnam before 1970).

It is exceedingly difficult to know whether KV is an accurate statement or not, and I doubt that there is anything approaching unanimity concerning it. But surely it isn't just a "matter of opinion."

Further, whether KV is true or not can be *rationally debated.* This point is crucial. One mark of a statement that is either true or false is that it makes sense to cite *reasons* for or against it. If I say:

(CB) "chocolate ice cream is better than vanilla,"

there isn't much I can do to defend my position. (In this context, "better" isn't intended to mean *nutritionally* better; for it can be debated whether chocolate ice cream is more nutritious than vanilla.) Of course, I can say "I *like* chocolate more than I like vanilla." But that doesn't provide very much support for my original statement.

KV is not like CB. For one can easily cite reasons for or against KV. One can say, "just prior to being assassinated, JKF wrote a speech that he intended to deliver to Congress explaining why he thought it urgent to pull out of Vietnam." Or one can say "recently uncovered documents show that, in LBJ's view, the United States would cease to be a world power if they pulled out of Vietnam; and those documents also show that he and JFK disagreed about this." And so on.

So, difficult though it may be to adjudicate, KV can be evaluated *in terms of* its agreement, or lack thereof, with objective facts (e.g., with documents,

video footage, people's memories, etc.).

If a statement's truth or falsehood is to be established *on the basis of* objective fact, that strongly suggests that the statement itself is one of fact, as opposed to "mere opinion." This is because, if a statement stands if the facts are one way and fails if they aren't that way, then that statement says, or at least implies, that the facts are that way. And, in that case, that statement has factual content.

Bottom line: any position that can be rationally disputed is likely to have an objective basis. KV can be rationally disputed, and that's why it's not a matter of "mere opinion." CB cannot, and that's why it is a matter of "mere opinion" or "preference."

15.2 Discussion of (ii)

Can ethical positions be rationally disputed? Yes. Let us suppose that A and B disagree about whether gay marriage is ethical. A would certainly make at least some case for gay marriage if he were to demonstrate that gay marriage, when it exists, is as likely to be based on love as heterosexual marriages. And B would undermine A's case and would also make a case of his own *against* gay marriage, if he were to show that, in fact, gay marriages are (or, if they were to exist, would be) less likely than straight marriages to be based on enduring feelings of love. Another way that B could make a case *against* gay marriage would be to show that, if it were condoned, the result would be a disintegration of practices that are necessary for the continued existence of any society. By the same token, if A could show that gay marriage would not have such a consequence, and that it might actually have the opposite consequence, then A would not only undermine B's case but would also make a case *for* gay marriage.

The point is that whether gay marriage is right or wrong depends largely on purely *factual* issues. The question "would gay marriage, if condoned, lead to the disintegration of social infrastructure?" is a strictly causal question. What it asks is whether a certain cause-effect relation exists. And the question "are gay marriages as likely to be based on enduring feelings of love as heterosexual marriages?" is a question of psychological fact; it is a question that would be of an interest to a psychologist who had no desire at

all to make value-judgments. In section (iv), we will provide two additional illustrations of the idea that ethical debates often turn on questions of fact and, for that reason, are themselves questions of fact.

15.3 Discussion of (iii)

Obviously, there are fierce, and also unending, disputes in ethics. Is abortion moral? Was it moral of the United States to go into Iraq? Is corporal punishment moral? These questions may be debated for centuries to come.

But there are many moral matters concerning which there isn't serious dispute. Is it morally okay to torture babies for the sheer fun of it? No. Is it morally obligatory to give all of one's money to charity? No. Is it praiseworthy (i.e., morally excellent) to give much of one's money to charity (when there is no practical incentive to do so)? Yes.

These questions may seem trivial. But, if so, that's precisely because it is so clear how they are to be answered. The illusion that there is *no* moral agreement—when, in fact, such agreement is less the exception than the rule—is that an ethical issue isn't worth discussing anymore if there is universal agreement about it. For this reason, the only ethical issues that are worth discussing are the ones that haven't been solved, giving rise to the misleading appearance that there is no ethical agreement.

15.4 Discussion of (iv)

Nonetheless, I wouldn't go so far as to deny that there is more disagreement in ethics than in most other disciplines. Nor would I deny that the basis of the disagreement is deeper. But, so far as that is true, it isn't because there are no ethical facts; it is because, in ethical discussions, people often shy away from making it 100% clear what, precisely, they are arguing for.

When we were discussing (i), we saw that if it is possible to cite reasons for or against a viewpoint, then that viewpoint, if correct, has an objective basis. If you know that Smith murdered Jones, then you are able to provide reasons why I should believe you. Maybe you can show me video footage of Smith doing the deed. Maybe you can remind me of how often Smith fumed about Jones. Maybe you can show me legal documents that make it clear that Smith stood to make a lot of money if Jones died.

By this reasoning, if the merits of ethical positions can be rationally disputed, then it is probable that ethical positions, when correct, have an objective basis. So can ethical positions be rationally debated? Can one cite reasons for and against them?

Yes. We saw this a moment ago when discussing marriage. Here is another example. Is the death penalty right or wrong? People disagree about this. But even among those who disagree, there is considerable debate as to what *would* have to hold *in order for* the death penalty to be justified. So, for example, if it were made clear that the death penalty systematically led to the execution of innocent people, then most people would agree that the death penalty was not right. The question "how many innocent people are executed for crimes they didn't commit?" is a strictly factual question. And a lot of the debate over the ethical integrity of death penalty depends on how this strictly factual question is answered.

Suppose it turned out that the death penalty made governments less likely to invest in social policies that would prevent the commission of capital offences in the first place. In that case, the vast majority of people who now support the death penalty as justified would probably cease to do so. (Of course, there would be a few holdouts. But that doesn't mean anything. There are still people who think the Earth is flat.) It is a strictly factual question of whether the death penalty inhibits, encourages, or is neutral with respect to the implementation of policies that discourage the commission of murder and other capital offenses. So, as before, we see that, to the extent that it's hard to decide whether the death penalty is good or bad, that's because certain purely *factual* issues are hard to decide (e.g., it's hard to know exactly what sociological effects the death penalty has). So the seemingly unending nature of this ethical debate has to do, not with there being no fact of the matter, but with there being relevant facts that aren't yet in evidence.

Another example: It's fiercely debated whether abortion is moral or immoral. But notice that the debate often turns on questions of fact. Does the fetus have a mind? Can it feel pain? Does it, at some level, feel an attachment to its mother and would it consequently feel an extreme sense of betrayal were it to find out that it was in the process of being aborted? These are factual questions about the fetus's psychological state (or lack thereof). And I think that many (though probably not all) pro-choicers, were they to find out

that fetuses clearly had minds would change their position. By the same token, many pro-lifers would change their minds were they to find out that fetuses clearly had no minds at all and were not otherwise comparable to *bona fide* human beings. When pro-lifers are asked to defend their view, they do so by citing (supposed) evidence that aborted fetuses have minds and, in particular, can feel both emotional and physical pain during the act of abortion. And when pro-choicers are asked to defend *their* position, they cite (supposed) evidence that fetuses don't have the neural wherewithal to have minds. So, while there is no denying that the debate over the merits of abortion has a more incessant, and also bitter, quality than debates about the structure of the hydrogen atom, much of the debate over abortion, and possibly all of it, does seem to turn on strictly factual questions.

16.0 What are ethical truths?

We're not in a position to answer this question right now. (An answer to this question is proposed and defended in Chapter 22.) But the points just made may point us in the right general direction.

We just saw that ethical contentions often have a heavy, albeit implicit, factual component. This suggests that ethical assertions *are* factual assertions, albeit disguised ones.

This view has been taken. According to one version of it, which is known as "utilitarianism," good things are those that bring about happiness and bad things are those that diminish happiness. (Not the happiness of any *one* person, but the happiness of *everybody.*) So a utilitarian would say universal health care is morally good if it increases happiness, morally bad if it decreases it, and morally neutral if it does neither.

We'll discuss utilitarianism in Chapter 24. But a couple of points about it are relevant in this context. First, it can't be all wrong. It's hard to believe that something that made everybody miserable would be good

or that something that made everybody happy would be bad. The question "does x make people happier, unhappier, or neither?" is a strictly factual question. Whether Smith is happy or not has nothing to do with my opinion or anyone else's. (It doesn't even have to do with Smith's opinion. His opinion is more likely to be accurate, than other people's, but if it's correct,

that's because it's rooted in the relevant facts. In fact, Smith's opinion as to whether he was happy *wouldn't* be more likely to be accurate than anyone else's if it were just a "matter of opinion" whether he were happy or not.) The question "how happy is so and so, if at all?" is a strictly factual question, and ethical judgments are not a "matter of opinion" if utilitarianism is correct.

According to another view, for something to be good is for it to be identical with, or logically required by, a practice on which human *flourishing* depends. By "human flourishing" is meant the actualization of human abilities. So, according to the view in question, if you have an innate talent for writing fiction, then you are flourishing insofar as you are putting that talent to use and failing to flourish insofar as you are not. It's a question of fact whether so and so is putting his innate musical ability to good use. Given any creature and any inborn ability on its part, it's a question of fact to what extent, if any, it's using that ability. And it's a question of fact to what extent a given act on your part or anybody else's promotes or thwarts the expression of inborn ability. So if the just-descried moral view is correct, moral questions are factual questions. "Is x good?" is identical with the question "does x promote or thwart the development of people's abilities?"[362]

Chapter 20

Emotivism

1.0 Are value-judgments expressions of emotion?

According to some[363], it is inherent in the nature of ethical pronouncements that, when made sincerely, they express feelings or emotions of some kind. Given only that you've sincerely said "that rock weighs over 100 pounds," nothing can be concluded about how you feel about the rock's weight. It's unknown whether you're happy about it, sad about it, or neither. But, so it is said, if you say "Larry is a bad man," your statement, if sincere, must embody feelings of some kind concerning Larry's deeds or character. If you feel nothing either way, you're guilty of insincerity in uttering that

sentence, and are in the same category as somebody who hates opera who says "I absolutely *adore* opera."

If correct, these points suggest that moral "judgments" are nothing more than expressions of sentiment. If you say "Larry is a bad man," you're just venting your negative feelings about Larry, and you're no more saying something true or false than you would be if you scowled at Larry. By the same logic, when you say "Mother Teresa was a good person," you are no more saying something true or false than you would be if you gave Mother Teresa a loving hug.

Thus, ethical statements express emotions, not judgments. "Stealing is wrong" is just a way of saying "boo stealing!" and saying "giving money to charity is good" is just a way of saying "hurray giving money to charity!" So-called moral "judgments" thus turn out to be nothing more than expressions of feeling and, therefore, to say nothing at all about the world. Consequently, there are no moral truths and no moral falsehoods.[364]

2.0 Why value-judgments aren't expressions of emotion

It's simply false that saying "x is good" is tantamount to saying "hurray x!" It's obvious that a competent speaker of English can sincerely say "Cotton Mather[365] was a good man" without personally having any emotion in connection with Cotton Mather. It's clear that people can *coolly* make moral judgments. In fact, an emotion-driven judgment is, to that extent, not an ethical judgment—or, in any case, not an unbiased one.

One often makes positive moral judgments about people and things that bring them displeasure. There are people who I dislike whose noble acts make me want to wretch. But, when not clouded by personal sentiment, I do judge that those acts are good. And when I say "in doing X, so and so did the right thing," I'm not saying "hurray X." In all likelihood, I'm wearing a frown when I grudgingly express this positive judgment.

But how is this to be reconciled with the emotivist's contention that ethical statements express emotion? David Hume (1739, 1751) provided a compelling answer to this question, which may be summarized as follows:

(HA[366]) The nature of moral judgment can be understood in terms of a fact about art. In general, people regard artistic works that bring them pleasure as “good” and they regard works that fail to bring pleasure, or bring displeasure, as “bad.” This suggests that, when said of a work of art, “x is good” means “x gives me pleasure,” and “x is bad” means “x fails to give me pleasure.” To be good, in the artistic sense, is to bring pleasure.

But this analysis cannot quite be right. People can, without self-contradiction, judge that works that don’t bring them pleasure are “good” and works that don’t do so, or bring them displeasure, are

"bad." This happens when people know that they *would* like a given work were it not for the inhibiting influence of some plainly irrelevant factor.

A story may help. Smith is a marvelous composer, and you know this. But you also detest Smith. One day you hear a performance of one of his works, and the experience is torture. But you know that, were it not for the fact that you don't like Smith, hearing that work would be exquisitely pleasurable. So you know that it's good, even though, because of this random fact about your past, it doesn't bring you pleasure. And you also know that, so far as other people's music appreciation isn't similarly blunted by such person-specific idiosyncrasies, they would derive pleasure from listening to this work of Smith's. In general, when people describe a work of art as "good," they mean that, setting aside irrelevancies of the sort just described, it brings people pleasure.

The same thing *mutatis mutandis* holds in the moral sphere. In general, people regard acts that bring them happiness as "moral" and regard acts that bring them unhappiness as "immoral." People act "immorally" when they beat you, and they act "morally" when they help you. But, of course, people sometimes regard things that bring them no pleasure, or even displeasure, as "moral"; and they sometimes regard things that bring them acute pleasure as "immoral." Many a person has committed heinous crimes and then commended the government for putting him in jail. In many such cases, if not all of them, the person is condemning acts that he obviously committed because doing so was itself enjoyable or was expected to bring him happiness down the line, and he is also commending an act on the government's part that will bring him untold misery.

But how is this possible, if, as we've seen reason to believe, for x to be good is for x to increase happiness? Because "x is good" means "x makes people happy to the extent so far as individual-specific, non-transferable facts about them aren't a factor." So the condemned prisoner who praises the government for putting him away knows that, were his circumstances not so abnormal, the government's putting him away would let him sleep easier at night (and its not doing

so would keep him awake in terror).

Hume's powerful argument shows that, if it is to be credible, emotivism must be changed; and it also makes it clear how to change it. Since, as Hume points out, a person can judge that something is good even if that thing brings him misery, it follows that, in direct opposition to what the emotivist holds, saying "x is good" cannot possibly be the same thing as expressing one's happiness about x. But "x is good" *is* plausibly seen as expressing the speaker's belief that x is happiness-conducive (i.e., that *ceteris paribus* x brings happiness). So, when modified along these lines, emotivism would be the view that "x is good" means "x is happiness-conducive."

But, thus modified, emotivism is blatantly false. If it expresses any belief, "x is good" is either true or false. A fortiori it's either true or false if it expresses the speaker's belief that x makes him happy to the extent that his idiosyncrasies aren't getting in the way. The essence of emotivism is that "x is good" expresses a feeling, as opposed to a judgment, and therefore doesn't say anything true or false. So emotivism collapses into its opposite if it's modified along the lines suggested by Hume.

Notice that, thus modified, emotivism coincides with utilitarianism. (This is the doctrine that an act is good to the extent that it maximizes happiness and bad to the extent that minimizes it. See Chapter 24.)

What is more, HA is viciously circular. If I hear a beautiful piece by Chopin and I judge that others *would* take pleasure in it, I am assuming that others, were they to hear it, weren't drunk when they did so, and that the parts of their brain responsible for musical enjoyment hadn't been destroyed. Any piece of music might fail to elicit a pleasurable response from any given person, *provided* that the latter were damaged in some such way. Thus, in assuming that others *would* like a given piece of music, what I'm assuming is that others would enjoy it *provided* that they weren't damaged—*provided,* in other words that they weren't for some reason incapable of seeing its merits.

"But saying that drunkenness or injury or some such made somebody incapable of enjoying the piece is different from saying that it made them incapable of seeing its merits."

Is it? Suppose that a blow to the head makes Smith incapable of enjoying music but doesn't otherwise change him. We'd say that Smith had been

damaged by the blow. But if his previous enjoyment of music had

no basis in the music itself—if it didn‘t derive from his being sensitive to it actual virtues—it would be wrong to say that Smith had been damaged in any way. On the contrary, we'd have to say that the blow had made him better. If his previous enjoyment of music had nothing to do with anything inherent in the music, his enjoyment of it was pure projection. But now, thanks to the blow, he doesn't respond to music in this delusional way. But since Smith obviously wasn't enhanced by the blow, and since he obviously was diminished by it, we have to say that his previous enjoyment of it did involve some awareness on his part of its merits.

It's crucial to distinguish between enjoying something, on the one hand, and being caused by it to experience pleasure, on the other. Given only that my hearing Chopin's ballad in G-minor gives me pleasure, it doesn't follow that I'm enjoying it. Maybe I know that my arch-nemesis Jones, who is in the room, hates it and, to my extreme delight, is in agony because it's blaring from the speakers. Here the G-minor ballad is causing me joy; but I'm not enjoying *it.*

There's a difference between being *caused* to feel good by X, on the one hand, and *enjoying* X, on the other. When this distinction is given its due, emotivism collapses, and so does every form of value-nihilism. (By "value-nihilism" I mean any doctrine to the effect that "x is good" either means nothing or means that x is liked or otherwise looked upon favorably by some person or group.) There are a lot of reasons why some object might make you feel good. Not all of those reasons are of an aesthetic nature.

A story will help make this clear. Smith owns a Picasso. He hangs it in his front hall, where visitors are bound to see it. The fact that it's there makes him feel good. It makes him feel good *not* because it's a great work of art, but because he believes that having it will improve his status and, therefore, his professional and romantic prospects. Smith neither likes nor dislikes the painting per se. In general, Smith neither likes nor dislikes art in general. He doesn't understand it. As far as he can tell, paintings are just discolorations on a canvas.

The Picasso makes Smith feel good. So his reaction to it is "hurray painting by Picasso that's hanging on my wall!" But this reaction isn't an *aesthetic* reaction. In other words, supposing that X is that painting, it isn't

X's being a work of art that makes him feel good. What makes him feel good is X's being something that will get him a promotion, and X's being something that will improve his social status, and so on.

Here a couple of points about causality may be helpful. No *thing* does any causal work. What *does* do causal work is some thing's having some property. The stovetop's being hot is what burns my hand; the stovetop per se does nothing.

Bearing this in mind, let us turn our attention back to Smith's painting, which we'll continue to call "X." X per se does nothing. X's having this or that property is what does the causal work. It is X's being impressive to employers that gets Smith the promotion. X per se does nothing. And supposing that Jones has a genuinely *aesthetic* reaction to X, what makes Jones feel good is X's having certain artistic merits.

An aesthetic reaction is one that is driven by a thing's *being a work of art.* In other words, it is one that is driven by a thing's having such and such aesthetic properties. So to the extent that there are aesthetic reactions, there are aesthetic facts—there are cases of things having such and such aesthetic properties. But once it's granted that there are aesthetic facts, the theory that "for x to be good is for me (or someone else) to like it" collapses. If there are aesthetic facts, we can be wrong about them, just as we can be wrong about anything else. If there are aesthetic facts, our feelings may or may not do them justice. So if there are aesthetic facts, value-nihilism is false, at least as far as aesthetic properties are concerned.

"But that's preposterous," it will be said. "Everyone knows the old saying: 'there's no disputing taste.' Your position embodies a failure to see the deep wisdom in that saying. In saying that there are aesthetic facts—that some paintings are better than others, that some musical works are 'more correct' than others, and so on—you are, in effect, saying that there are paintings, works of music, novels, and so on, that people *should* like. In addition to being ludicrous, that view is potentially very dangerous."

This viewpoint embodies several confusions. First of all, given only that there are aesthetic facts, it doesn't follow that, given *any* two works of musical works, it doesn't follow that one is better than the other or that they're equally meritorious; in other words, it doesn't follow that they are

capable of being meaningfully compared. (In this context, take my use of the term “musical work” to be short-hand for “musical work *or* novel *or* painting, etc.”) Musical works have to be understood on their own terms. Beethoven’s 24th piano sonata (in F# major) is good; so is Tchaikovsky’s violin concerto; so is “Dark Side of the Moon,” by Pink Floyd.

But given any two of those musical works, they are good in non-comparable ways. (The goodness of any one of them is "incommensurable," as some ethicists put it, with the goodness of either of the others.)

That doesn't mean that every musical work is as good as every other musical work. "Dark Side of the Moon" isn't *supposed* to be good in the same way as Beethoven's 24th Sonata; and that sonata is *supposed* to be good in the same way as Tchaikovsky's violin concerto. But there *are* works that *are* supposed to be good in the same way as Beethoven's 24th sonata but, at the same time, *aren't* good in that way, at least not to the same degree. (Any one of Clementi's piano sonatas would be an example.) There *are* musical works that, given the genre to which they belong, are supposed to be good in the same way as "Dark Side of the Moon", but *aren't* good in that way, at least to the same degree. (An example would be "Beast of Burden" by the Rolling Stones, which even the Rolling Stones themselves would admit isn't one of their better songs and isn't as good as "Dark Side of the Moon.")

In any case, it's patently obvious that some composers are better than others. I used to compose. I was better at it than the other people in my high school. But when I moved beyond the pathetically narrow horizons of my high school existence, I found that I was seriously outclassed. I wish I could compose like Mozart (or Roger Daltry); but I cannot. Many a musician wishes that he were a better composer than he is. And no such musician can coherently believe that no musical work is better than any other. For if all musical works were equal merit, then *ex hypothesi* all composers would be of equal merit.

Moral nihilism fails for much the same reasons that aesthetic nihilism fails. I see wealthy person X give money to poor person Y. Let's refer to this situation as S. S makes me feel good. It makes me feel so good, in fact, that I go so far as to say "hurray S!" Given only that S makes me feel good, it doesn't follow that I'm having a *moral* reaction to S. Maybe I just happened to like the way that the crisp green of X's money contrasts with the cool blue of the sky. Maybe some fact about X's posture shows me that some anatomical theory of mine is correct. Given only that I *like* S, it doesn't follow that I'm having a *moral* response to it. It's only if S's having such and such *moral* properties is what brings me joy that my liking S is a moral reaction to it. So my exclaiming "hurray S!" expresses a *moral* reaction to S

only to the extent that S's being a moral fact is what evokes the emotion thereby expressed. In general, there are moral reactions only to the extent that there are moral facts. If there are moral facts, our feelings may or may not do them justice; and emotivism completely collapses.

But we have yet to state the most basic problem with emotivism.

3.0 The most basic problem with emotivism: the doxastic nature of emotion

Our sense-perceptions represent the world as being a certain way. Some are accurate; some are not. But, whether accurate or not, sense-perceptions always depict or represent the world as being a certain way; they always "say something about it." By this I mean only that, like a spoken or written statement, they have "truth-conditions": they're accurate ***iff*** the world satisfies certain conditions.

The same thing is true of thought. Some of our thoughts are true; some are false; some are a mixture of truth and falsehood. But, whether true or false, our thoughts always represent things as being a certain way; they always say something about the world.

Do *emotions* say anything about the world? Are they, in this respect, like perceptions and beliefs? "Surely not!" it is said:

> Emotions can be healthy or unhealthy, but they cannot be true or false. It's not as though, by virtue of being happy or sad, you're right or wrong about anything. You can be happy that Tim left home, sad about it, or neither. But none of these feelings is any more accurate than any of the others, even though some may be more pleasant or more adaptive. In general, emotions are neither true nor false, neither accurate nor inaccurate.
>
> Of course, emotions typically involve presuppositions about how the world is. You cannot be angry *that Tim stole your bicycle* unless you believe that Tim stole your bicycle. You cannot be happy *that Sally got married* unless you believe that *Sally got married.* And, of course, your belief that Tim stole your bicycle is either true or false, as is your belief that Sally got married. But your *anger* at Tim's (supposed) act of theft isn't true or false, and neither is your happiness

about Sally's nuptials.

> Emotions *presuppose* beliefs; but emotions aren't themselves beliefs. If you're happy, you're happy *that such and such is the case.* So you're happiness presupposes that you believe that such and such is the case. But, while that belief is either true or false, your happiness about it is neither. The same is true of all other emotions. If you're worried, you're worried that such and such might happen and you therefore believe it possible that such and such might happen. This *belief* may or may not be correct. (Your worry is "well-grounded" if that belief is correct, and it's otherwise not well-grounded. But your worry itself is not true or false.) So, whatever it is that your being worried *adds* to your belief, it isn't something true or false. What it adds may be important from a personal viewpoint; but it's neither here nor there as far as your knowledge of the world is concerned.

The position just described was advocated by David Hume, and we'll henceforth refer to it as "Hume's view." Hume's view, we will now see, is totally false. Emotions *are* true or false. The reason is that, counterintuitive though it may seem, *emotions are beliefs.*

If you have a headache, you'll happily take a pill to get rid of it. We don't see it as strange for someone to take a pill to get rid of an unpleasant *feeling.* We don't feel that they're self-lobotomizing. But that *is* how we'd regard somebody who took a pill that wiped away all his unpleasant *emotions.*

Your child has been kidnapped. You suspect that he's still alive. You don't know who the kidnapper is or where he's keeping your child, and neither do the police. The police are not pursuing obvious leads and seem not to be doing much to find your child. You are angry at them, but you're 10 times angrier at the kidnapper.

By sheer coincidence, a friend of yours—a physician and world-renowned biochemist—has invented a pill you can take that will get rid of your anger. The pill has no side effects and it won't cost you anything.

Would you take it? Probably not. Would your attitude towards taking it be comparable to your attitude towards taking an aspirin to get rid of a headache? Definitely not. What you want isn't to get rid of the emotion.

What you want is for the police to find your child. It's true that, if they find him, your anger will vanish or assume other, milder forms. But it isn't *because* you want to get rid of your anger that you want the police to find your child. For if it were, then their finding the child would just be a means to that end, in which case you'd just as soon take the pill as you would have them find your child. But that isn't what you'd do. If you had to choose between taking the pill and having the police find your child, you would, without any doubt, choose the latter. You'd even be willing to endure an intensification of your anger if you thought it might make it more likely that your child would be found. So your desire to find your child is *not* a desire to make your anger go away.

In general, what people want isn't to extinguish the unpleasant emotions they have. What they want is to change their circumstances in such a way that their unpleasant emotions no longer have a basis in them. People don't want to "zap" their anger or hate or pessimism. They want the world to change in such a way that those emotions cease to be appropriate to it.

In this respect, the attitude that people have towards their emotions is similar to the attitude they have towards their senses. People want to see beauty, not ugliness. But they want to see beauty *because it's really there*—not because they took a pill that made them see a pristine forest-stream where there was actually a rivulet of septic gutter water. And people want to feel joy. But they want to feel it because their circumstances warrant it—not because they took a pill that tricked them into feeling joyous in sorrow-warranting circumstances.

Like a sense-perception or thought, an emotion must have an object. You can't *just* be angry. You're angry at your boss *for* giving someone else the promotion. You're angry at Tim for stealing your bicycle. Your anger is therefore *about* something—it's about your boss's snubbing you or Tim's robbing you.

The same thing *mutatis mutandis* is true of all emotions. You're not *just* remorseful; you're remorseful *that* you sold your wife's jewelry to buy drugs. And you're not *just* relieved; you're relieved *that* she's not pressing charges.

3.1 The contentfulness of emotions *qua* emotions

As previously stated, emotions have propositional complements. You're not just angry; you're angry *that* Tim stole your bicycle. Your anger at Tim isn't *identical* with your belief that he stole your bicycle. After all, you could have that belief without being angry at him. At the same time, you're being angry at him for doing this necessarily involves your having that belief. So one component of your anger at Tim is your belief that he stole your bicycle. But there's some second component to your anger. What is it?

Since, of course, it's this second component that separates simply believing something from emoting about it, it's tempting to think that this second component, unlike the first, isn't true or false and, therefore, that it is nothing more than a *feeling* of some kind. So, according to this view, your anger at Tim consists in (1) your belief that he stole your bicycle plus (2) a negative feeling associated with his doing so.

But how is the second thing "associated with" the first? First of all, the mode of association cannot be purely *causal.* If, because of some anatomical quirk of yours, your belief that you had decent and loving parents caused you to have gastric distress, it wouldn't follow that you were *angry that your parents were loving ones.*

To generalize this point, given only that x is what *caused* so and so to be distressed, it doesn't follow that x is what so and so is distressed *about.* Even more generally, given only that x is what *caused* so and so to have emotion y, it doesn't follow that x is what, by virtue of having y, so and so is emoting *about.* The statement:

(A) x caused so and so to have emotion y

doesn't say the same thing as

(B) in having emotion y, so and so is emoting about x.

3.2 What emotions are

Taken in conjunction with the last point, a story from my own adolescence will help make it clear what emotions actually are. When I played soccer in high school, one of my team-mates, who we'll call "Ben," had a rather serious respiratory disorder, which caused his lungs to fill up with fluid whenever he did vigorous exercise for more than a few minutes. But Ben was

an ardent soccer player, and he played as much as his condition would permit. During a match, he'd play for as long as he could, and then he'd have to take a time out to let his lungs drain, during which time he experienced painful, wracking coughs.

During these time-outs, Ben was upset. He was upset about his symptoms, about his inability to breathe properly, etc. Of course, it was playing soccer that set off these symptoms. But he wasn't upset *that he had played soccer.* That was the one thing he was happy about. This confirms that having a negative emotion *about* X isn't the same thing as being caused by X to have a negative emotion. What upset Ben was that his doing something which *didn't* upset him caused him such unpleasantness. Another thing he was upset about was that his illness prevented him from playing soccer as much as he wanted to.

Ben would obviously be happy if he were cured of his illness and could therefore play soccer without any problems. But he wouldn't *just* be happy. He'd be happy *that he could play soccer without any problems;* he'd be happy *that he was no longer prevented from doing what he wanted to do.*

I therefore propose that *Ben's being happy under that circumstance would be identical with his recognizing that he could do something he wanted to do.*

Consider an alternative universe W where Ben is exactly as he is here, *except* that he doesn't have any desire to play soccer (or to do or have anything for which playing soccer is a necessary prerequisite). In W, Ben has the earlier-described malady and he's unhappy about it, but he isn't *unhappy that it prevents him from playing soccer.* In W, he never wanted to play soccer in the first place. In W, there's no desire that he has that, in not being able to play soccer, he cannot gratify. So when he finds out that he's cured and, therefore, can now play soccer, he won't realize that he can now gratify a previously frustrated desire to play soccer. The reason he won't realize it is that there's nothing to realize, since he doesn't have that desire to begin with. Thus, a necessary condition for Ben's

being happy that he can play soccer is that he knows himself to have some desire that, in now being able to play soccer, he is able to gratify.[367]

It's also a *sufficient condition.* To see this, let's turn our thoughts back to the real world, where, sadly, Ben has an ardent desire to play soccer that his illness prevents him from gratifying. If, in actuality, Ben were suddenly cured of his illness, and he thus realized he could fulfill his life's dream of playing soccer, could he possibly *not* be happy? No, and it would be incoherent to say otherwise.

Thus, a necessary and sufficient condition for Ben's being happy that he can now play soccer is he knows himself to have some desire that, in now being able to play soccer, he is able to gratify. Ben's being happy that he can play soccer *is* his recognition that, in now being able to play soccer, he can now gratify an otherwise ungratifiable desire.

3.3 What emotions are (continued)

This analysis is easily extended to apply to all emotions. Another story will help us do this. You've invited your father to come visit you. But there's a chance that your malignant, sociopathic step-mother (his second wife) may come along with him. (If she comes, you won't kick her out, since doing so will alienate your father, which you don't want to do.) You thus fear that she will come.

Why is her coming bad? Why do you fear it? You wouldn't fear her coming if there were nothing that you wanted that, so you believe, her coming would be likely to prevent. Imagine that you were totally convinced that her coming wouldn't jeopardize anything that you hoped to happen—that it wouldn't prevent you from having bonding moments with your father, or having a beer with him while talking about the good old days, etc. Under that circumstance, could you *fear that* your step-mother would come?

No. You could believe it, but not fear it. This is because to fear her coming *is* to believe that something you want is jeopardized by the, in your view, distinct possibility of her coming. So, to fear her coming *is* for you to believe that her coming would frustrate some desire of yours.

Of course, you may be wrong about this; it could be that she does come and everything turns out fine. But that's consistent with our point that your fearing her possible visit involves your believing, not necessarily correctly, that her coming would make something you want impossible and would thus

thwart a desire of yours. And it actually bears out our main point, this being: emotions *are* true or false, since emotions are beliefs of sorts.

Emotions, then, are beliefs about the prospects of gratifying desires that one has. If the belief is that the prospects are good, the emotion is a positive one (e.g., jubilation, relief). If the belief is that those prospects are bad, it's a negative one (e.g., anger, remorse).

4.0 Geach's anti-emotivist argument

Peter Geach put forth the following, very clever argument against emotivism ("GA" stands for "Geach's argument"):

> (GA) Suppose that emotivism is correct. Suppose that saying "stealing is wrong" is, as far as its content is concerned, no different from saying "boo stealing!" and that saying "giving money to the needy is good" is the same as saying "hurray giving money to charity!" In that case:
>
> (1) "If *stealing is wrong,* then getting your little brother to steal is wrong"
>
> has the same meaning (or, if you prefer, is characterized by the same lack of meaning) as:
>
> (2) "If *boo stealing!,* then getting your little brother to steal is wrong."
>
> Here's the idea. (2) is just like (1), except that (2) contains the expression "boo stealing!" in the place where (1) contains the expression "stealing is wrong." The emotivist says the underlined part of (1)

has the same meaning (or lack thereof) as the underlined part of (2). Replacing an expression with a synonymous expression doesn't turn meaning into nonsense; replacing the occurrence of "enemy" in "Smith is Brown's enemy" with the word "foe" results in "Smith is Brown's foe," which is as meaningful as the first sentence and, in fact, has the very same meaning as it. But whereas (1) is obviously meaningful—not necessarily *true,* but meaningful—(2) is complete nonsense. Thus, the underlined expressions are not synonymous, and the emotivist is wrong to say otherwise. Given any emotivist "translation" of any moral statement, an exactly analogous argument shows that translation to be wrong.

4.1 Evaluating Geach's argument

I've argued that emotions *are* beliefs. Let's suppose that, after 10 years of your being an unremittingly good friend to him, Smith turns his back on you. He does this to curry favor with his boss. You're outraged and appalled; your contempt for Smith is boundless. These emotions, I have argued, *are* beliefs. Like all emotions, they are, I have argued, beliefs about desires of yours. Your outrage, for example, might be identical with your belief that Smith's behavior thwarted your desire for a relationship in which certain rules of decency were observed. For expository reasons, let's suppose this to be the case.

Let's also suppose that your outrage is initially expressed by your throwing a chair across the room while yelling "aarrr!" Finally, let us suppose, if only for argument's sake, that my analysis of emotion is correct. I wish to make it clear that, given this supposition, your yelling "AAARR!" while throwing a chair across the room expresses the aforementioned belief.

Question: Given all of these suppositions, does it follow that, when they are uttered by you, the following two sentences have the same meaning?

(3) "If *I believe that Smith's behavior thwarted my desire for a relationship in which certain rules were observed,* then I believe that Smith's behavior thwarted at least one desire of mine."

(4) "If AAAARRR! [accompanied by you throwing a chair across the room],

then I believe that Smith's behavior thwarted at least one desire of mine."

No it does not. A belief can be expressed in a lot of different ways. Your belief that 1 + 1 = 2 can be expressed by your choosing to buy certain stocks as opposed to others, by your deciding not to jog an extra mile in the rain, by your saying "1 + 1 = 2," by your pounding your chest with pride (after it's announced on television that your staggering mathematical insight was verified by M.I.T. mathematicians), and in any number of other ways. Similarly, your belief that Smith's conduct thwarted your desire can be expressed by your uttering certain sentences (e.g., the underlined part of (1)), or by your throwing chairs and bellowing.

Any given attitudes can be expressed in many different ways. One's love for another is expressed in many ways; and the same is true of anything else that one might feel for a person or thing—hatred, admiration, contempt, fear, envy, respect, and so on. If, as I'm proposing, emotions are beliefs, they can be expressed through grammatical utterances and meaningless gesticulations and facial contortions. Which makes it obvious that Geach's argument fails if emotions are beliefs. Which, in its turn, shows that it makes a strong, unargued, and probably false assumption. Which, finally, shows that it's spurious.

Chapter 21

Moral Conventionalism and Moral Nihilism

In this chapter, we will consider six arguments for the position that moral truths either don't exist or do exist but only in some degraded sense.

> ***Argument #1:*** It's preposterous to suppose that there were any moral laws before there were people. Therefore, such laws, if they exist, are human creations—just like the laws created by a nation's legislature. A nation's laws are for its legislature to create and modify. Similarly, moral laws are for us to create and modify.
>
> Whether an act is good or bad depends on what the laws of morality are. Thus, given that we create those laws, we decide what is good and what is bad. If polygamy is bad, it's because we created a moral law prohibiting it. If it isn't bad, it's because we didn't create such a law. It is therefore a matter of convention, not of fact, that keeping promises is good and having multiple spouses is bad.

Analysis of argument #1 For argument's sake, let's suppose that, just as the argument says, there were no moral laws before there were human beings (or intelligent beings of some other kind). It doesn't follow that human beings *created* them; nor does it follow that morality is in any way a matter of opinion. There are biological (anatomical, physiological) truths about people; and there were obviously no such truths before there were people. But biological truth isn't "relative" or "subjective" or anything of the sort.

But the parallel between morality and human biology goes much deeper than this. We come into this world with biological and psychological structures that, although capable of being changed within certain narrow limits, are otherwise quite fixed. These structures being what they are, we are easily damaged and we also need certain things to flourish. Morality is to be understood in terms of these needs of ours: it is fixed by them. What damages us is immoral, and so is what prevents us from flourishing. What heals us is moral, and so is what facilitates our flourishing.

Moral truths ***supervene*** on psychological and biological truths. Two situations that are identical in respect of their non-moral (descriptive)

characteristics cannot possibly differ in respect of their moral characteristics. [368] Once the physical, physiological, and psychological facts are fixed, so are the moral facts. This strongly suggests that *are* psychological and biological truths—that they are truths to the effect that certain things will damage people and others will augment them. (So far as it's moral to damage people, it's *only* to the extent that doing so is necessary to prevent them from being damaged in some more significant way.)

There are, of course, different views as to what constitutes morality.

Given the fact that, at their core, human biology and psychology are relatively invariant, the fact that there are many different ethical systems calls into question my contention that morality is fixed by facts about our biology and psychology. And the fact that many of these moral codes are punitive, both to their own adherents *and* to those who are on the receiving end of the conduct of their adherents, casts further doubt on the viability of my contention.

But given only that there different views as to what is moral, it doesn't follow that all of them are right. Some of them seem quite wrong. The Nazis had a moral code. It was a bad one; the ethical precepts it embodied were false. Many other moral codes are bad and, therefore, embody false ethical precepts. As for the fact that many moral codes are punitive (in the double-edged way just described), that is easily reconciled with my thesis. Left to his own devices, the individual behaves in ways that are radically anti-social. Therefore, a non-restrictive, non-punitive moral code is one that gives people no protection against their brethren and that therefore fails of its purpose. As for the fact that some moral codes are *more* punitive than they have to be, that is easily explained in terms of the fact that moral censure is one of the outlets for aggression and antipathy that doesn't jeopardize the social order.

We have both moral and also strategic obligations to *tolerate* different moral views. Given only that individual (or nation or culture) X has an excellent moral code and Y has a bad moral code, it doesn't follow that X can force his moral code on Y; it doesn't follow that X has the right to intercede in Y's affairs. For it may be that the damage done to people by intervening in their affairs is greater than the damage that such an intervention would prevent. If we launch a wholesale invasion on a country, so as to prevent its otherwise superb president from being mean to his children, we'll do more harm than good; and what we're doing is immoral, even if, thanks to it, the president's children are never again mistreated.

There's another reason why it can be immoral to be intolerant of what is immoral. People can't flourish without having a certain amount of freedom. Therefore, it is wrong to inhibit other people's freedom (except in so far as such abridgments are necessary to prevent other, greater ones). Were X to force its moral views on Y, the resulting abridgment of Y's freedom might be more immoral than the misbehavior on Y's part that it was X's intention to stop. And in that case, X's intervention would be immoral, even though Y's moral code was itself immoral.

Moral principles as conditional truths Thus, far we've conceded to Argument #1 that moral truths don't pre-exist human beings. But there's no need to make that concession.

Let "PBS" be our short-hand way of referring to our psychological-biological structure. Moral principles can be seen as being to the effect that *if* there were creatures having PBS, *then* it would follow that treating those creatures in certain ways would be wrong while treating them in certain other ways would be right.

Suppose that strange evolutionary turn creates creatures that, although not biologically human, are friendly and are capable of connecting with human beings in an emotionally meaningful way (e.g., they would care about people, come to trust them, etc.). Even, so far as we know, though such creatures do not *now* exist, we know that it *would* be wrong to cause them unnecessary suffering; we know that it *would* be wrong to alter the condition of such a creature in a painful or damaging way.

By the same token, enslaving human beings was wrong *even before there*

were people. The principle that enslaving another person is wrong. That principle can be seen to the effect that, *if* there are entities with PBS, they shouldn't be enslaved.

Hypothetical truths hold of whether or not either the antecedent or the consequent is true. Even if Smith has no money, it's true that "*if* Smith has a million dollars, *then* he has more than twenty dollars." So while it's true that, before there were people, the antecedent of (e.g.) '*if* there are people, they shouldn't enslave one another' was false, it doesn't follow that the whole hypothetical was false.

It may seem strange to say that, before there were people, there would be truths, even hypothetical ones, about them. But this feeling is totally unfounded. Many hypothetical truths concern things that not only don't exist but, for reasons of physical law, couldn't possibly exist, for reasons of physical law. Consider the statement:

"Euclidean triangles have interior angles adding up to 180°." Even though it doesn't look like one, that statement is a hypothetical, since it's meaning is: *if* x is a Euclidean triangle, *then* x has interior angles adding up to 180°.[369] The existence of Euclidean triangles is physically impossible; the structure of space prohibits it.[370] Human beings are, quite obviously, physically possible; so it's not clear why, in a person-free world, principles like "if there are people, they shouldn't be enslaved" wouldn't be correct.

To see exactly why Argument #1 is wrong, it is necessary to make two distinctions, namely: (i) the distinction between a moral law's existing, on the one hand, and its being known, on the other; and (ii) the distinction between moral law's being known from its being implemented.

Suppose that, at some point in time, it wasn't felt that slavery was wrong; and further suppose that, at the time, slaves were so damaged by their condition that they felt it to be a perfectly just one. Does it follow from these suppositions that slavery wasn't wrong? Is one person's brutalization of another morally right just because both parties happen to accept some spurious rationalization of it? Surely not. This entails that a moral law can exist *without* being known. Since we'd always know the laws of morality if we created them, it follows that we don't create them.

Realistically, slave-owners probably *did* know that slavery was wrong. (They may not have *said* it was wrong. But so what? People lie all the time, especially if it's in their economic self-interest to do so.) If people in a given society believe themselves to derive enormous benefits from a practice, they're not likely to stop it, even if they see it as deeply immoral. Slave-owners benefit from the institution of slavery. So they have no

reason to condemn it. They thus have a strong incentive to construct spurious rationalizations as to why slavery isn't always wrong. If, while continuing to have slaves, slave-owners brazenly admitted that slavery was evil they would mobilize otherwise inert abolitionist forces. If they don't admit this, and instead rationalize what they're doing, a thin but powerful margin of doubt will keep some of these forces at bay, and will also enable themselves to live with themselves.

It's often convenient for people to turn a blind eye to moral truths. For this reason, moral laws, even when known, aren't always *implemented* and also aren't explicitly endorsed.

It follows that moral laws aren't conventions. Conventions cannot fail to be implemented. It cannot be a convention that people wear green hats on Sunday unless people *actually* do so.

> ***Argument #2:*** *The argument from cultural relativism.* Moral convictions vary from culture to culture. In some cultures, polygamy is accepted. In others, it isn't.
>
> One possibility is that, given an ethical divergence between two cultures, one of those cultures must be wrong. So, just as two people can't disagree about whether Helsinki is the capital of Finland without at least one of those people being wrong, so two cultures can't have different ethical viewpoints without at least one of them being wrong.
>
> Another possibility is that, when two cultures disagree in this way, neither is right and neither is wrong. One culture likes polygamy. The other doesn't. And that's the end of it.
>
> The first option is implausible. If I like tennis, and you prefer soccer, it makes no sense to say that one of us is right and the other is wrong. If culture X likes polygamy and Y does not, it makes equally little sense to say that one of them is right and the other wrong.

So the second option must be the right one. When cultures diverge ethically, neither is right and neither is wrong. But if there were moral *facts,* somebody would be right and somebody would be wrong. Therefore, there are no moral facts.

Analysis of Argument #2 First of all, people knowingly do all sorts of immoral things. So given only that such and such is done in a given culture, it doesn't follow that people in that culture think that it's right to do such and such.

Second, people who know that what they're doing is wrong don't always admit it. In fact, they're exceptionally likely to lie about it. So given only that two cultures diverge in what they *claim* to regard as moral, it doesn't follow that there's any *actual* divergence of moral outlook. Given only that, in some culture, it's officially legitimate to own people of a certain race or gender, it doesn't follow that people in that culture generally think it's good. It could just mean that the people who benefit from this custom have the power to keep it going and believe (correctly, no doubt) that it'll be easier for them to do so if, they provide some specious ethical basis for it. If, instead, they brazenly admit that what they're doing is wrong but *still* insist on doing it, they'll lose many would-be supporters and will soon have to forego the practice in question.

Also, sometimes apparent moral disagreements are disagreements as to how to implement a shared moral value. If people in culture X think that beating children prevents them from becoming effective adults, whereas people in Y think that beating is necessary to make them become effective adults, then the moral value (*you should help your children become effective adults*) that underlies the acceptance of child-beating in X coincides with the value that underlies the prohibition of it in Y.

Having said all this, let's suppose that sometimes different cultures really do have different values. Does it follow that no cultures are ever, in any respects at all, morally more in the right than others? It does not, and it would be incoherent to say otherwise. Culture X tries to wipe out Culture Y. X has no good reason for doing this. (Y isn't a threat to X. Invading Y won't stimulate X's economy. And so on.)What is a cultural relativist to say? If he says that it's wrong, he's conceding that there is an intercultural value-system

and is thus undermining his own position. If he denies that it's wrong, he's saying nothing that one culture does to another is wrong. If correct, this position entails that there's nothing wrong with obliterating a culture and its value system with it. And this, in its turn, entails that a culture's morality is no more valid than its army is powerful. In other words, might makes right. But this is the antithesis of the sentiment of tolerance that (supposedly) underlies acceptance of cultural relativism. Thus, cultural relativism is self-defeating and therefore false.[371]

What they do in culture X is wrong ≠ we have the right to stop them from doing it

Given only that some culture engages in a practice that we know to be wrong, it doesn't follow that we have the right to intercede. First of all, the cure can be worse than the disease. The Dictator of Country X, let us assume, beats his wife. This is immoral. But if we were to invade X to prevent this, the result could be a breakdown in order that was followed by a flood of murders, rapes, etc., which, collectively, would make for a situation far worse than the one we were trying to correct.

When we make predictions about a country, we tend to rely heavily on knowledge of regularity-conducive structures, such as governments and other stable institutions. When a country is invaded, those institutions are *ex hypothesi* undermined. Therefore, our methods for predicting occurrences in that county cease to be reliable. An intervention isn't justified unless there's a reasonable expectation that the post-intervention situation is better than the pre-intervention one. Since interventions make it hard to have such expectations, it follows that they are often unjustified.

So it's one thing to say that *we* have the right to change X if female circumcision is common practice there. It's a very different thing to say that it is *right* of X to countenance this abominable practice. A failure to see that these are entirely different propositions underlies the view that whatever goes on within a nation's (or culture's) borders is *ipso facto* morally good.

Another relevant point is that no person and no nation would have much autonomy if it weren't allowed to do anything wrong. Nobody would have much freedom under a government that micromanaged individual morality, and no nation would have much freedom under an international government

that micromanaged national morality. It's obvious that autonomy is a good thing. Since a precondition for personal or national autonomy is being allowed, within limits, to do wrong, it follows that one can have an obligation to *refrain* from preventing acts that one *knows* to be wrong. This shows the spuriousness of the inference from "what so and so is doing is wrong" to "we have the right to stop so and so from doing it."

> ***Argument #3*** Science has no use for moral facts. No accurate scientific description of the world, no matter how exhaustive, would ever mention them. It would mention atoms and protons, but not instances of wickedness or decency; and, while it would describe things as having positive or negative charges, it wouldn't describe anything as being right or wrong. Since moral facts have no place in any scientific description of reality, we have no reason to grant their existence and every reason to deny it.

Analysis of Argument #3 If ethical facts *are* natural facts, then science *does* need to grant their existence. According to ***utilitarianism,*** an act is good to the extent that it promotes happiness and bad to the extent that it promotes unhappiness. According ***virtue theory,*** an act is good to the extent that it promotes the development of inborn potential and bad to the extent that it prevents it.[372] Science doesn't deny that some acts do more than others in the way of creating happiness or encouraging the development of inborn potential. So if these doctrines are right, moral facts *are* natural facts and, therefore, must be countenanced by science. Utilitarianism identifies goodness with the property of being happiness conducive and, therefore, with something whose instances fall within the scope of natural science. The same thing *mutatis mutandis* is true of virtue theory and, indeed, of most ethical theories.

Those who advocate #4 seem to believe that, if two sentences describe some one fact, they must be synonymous. They seem to think that if "x is good" registered the same fact as some sentence belonging to physics or biology, it would be *synonymous* with such a sentence.[373] That isn't true. The fact registered by "Smith has two cars" is identical with the fact registered by "if n is a number such that n = 2 iff arithmetic is incomplete and

such that n = 3 otherwise, it follows that Smith has n cars." But those sentences aren't synonymous.

> ***Argument #4*** "Abortion is morally wrong" means "abortions shouldn't occur'; and "keeping promises is morally right" means "promises should be kept." "Smith is wrong to beat his children" means "in beating his children, Smith is doing something he shouldn't do"; and "Smith is right to tend to his ailing mother" means "in tending to his ailing mother, Smith is acting as he should." In general, setting aside a few nuances, "x is good" means "x should exist (or occur)" and "x is bad" means "x should not exist (or occur)."
>
> So far as a statement fails to say that something should or should not be the case, it isn't an ethical statement. "Smith is tending to his ailing mother" isn't an ethical statement, since it doesn't say that anything should or should not occur. So far as a statement *does* say that something should or should not occur, it *is* an ethical statement. "Smith is right to tend to his ailing mother" is an ethical statement, since it says that, in tending to his mother, Smith is acting as he should. Thus, ethical statements say what *should* occur, not what *does* occur.
>
> But reality is described by saying what there is, not what there should be. If true, "Smith killed Jones" describes reality; it says what happened. So far as "Smith was wrong to kill Jones" describes reality, it says the same thing as "Smith killed Jones." So far as it doesn't say the same thing, it doesn't describe reality and therefore says nothing. So "Smith was wrong to kill Jones" says nothing not also said by "Smith killed Jones." In general, "x exists and x should (or should not) exist" says nothing not said by "x exists." Therefore "x exists and is good" says nothing not said by "x exists," and neither does "x exists and x is bad." The words "x is bad" and "x is good" add nothing; so they say nothing. And since they say nothing, they say nothing true or false. "x is good" isn't true or false, and neither is "x is bad." Ethical "statements" aren't true or false and therefore aren't even statements. They're pseudo-statements, not real ones.

Analysis of Argument #4[374] Moral imperatives are ***hypothetical*** imperatives. A hypothetical imperative is one of the form: *given* that you

want x, you should do y. Hypothetical imperatives always fall within the scope of natural science, as we will now see.

You cannot give an amoral person any good reason to be moral. Given that he is amoral and that his values radiate from that fact, it would be deeply irrational of him to heed your injunction to be moral. Of course, it's sometimes to one's practical advantage for one's *behavior* to coincide with that of a moral person's. But that's irrelevant, since an act done for prudential reasons isn't a moral act, even if it's externally indistinguishable from one.

Any reason that anyone has to be moral is thus of a purely *hypothetical* nature and is expressed by a sentence of the form:

(i) *Given* that your objective is to be moral, you should do X.

Different ethicists have different views as to what X is. According to utilitarians, X is what maximizes happiness. If utilitarianism is right, then (i) means:

(ii) Given that your objective is to maximize happiness, you should do X.

(ii) is just a garden variety causal statement, since it's to the effect that:

(iii) Not doing X will have the consequence that people aren't as happy as they would otherwise be.

Thus, the "shoulds" of ethics are like the shoulds of engineering. The statement:

(iv) *Given* that your objective is to construct a sturdy bridge, you should use material X

is a way of making the purely causal claim that:

(v) Not using material X will have the consequence that the bridge isn't as sturdy as it would otherwise be.

Many ethicists reject utilitarianism. But given practically any ethical theory, a consequence of its being correct is that its moral imperatives become garden-variety causal statements, and thus fall within the scope of natural science.

Argument #5 Ethical statements have a property that sets them apart from all others: if understood and believed true, ethical statements *necessarily*

incline one towards certain actions. If you really understand and agree with the principle that you should keep promises, you cannot fail to have an inclination to keep your promises.

This doesn't mean that you necessarily *will* keep your promises. You may have other inclinations that override your inclination to keep your promises. For example, given your burning wish to be a millionaire by age 40, along with your correct belief that this wish of yours won't come true if you're utterly scrupulous, you may choose not to keep your promises, even though you wish that circumstances didn't force you to break them. But if you have *no* inclination to keep your promises, then you either don't understand the principle that you should do so or you don't really agree with it.

So far as this appears false, it's because it's easy for people who don't understand or agree with ethical principles to impersonate those who do. Somebody who, while not having a shred of ethical insight, knows the right talking points can easily pass for somebody who not only understands, but also accepts, the precept that one should keep one's promises. Be that as it may, what is relevant here is that wherever there's actual understanding and acceptance of an ethical principle, there is, for that very reason, an inclination to act in accordance with that principle.

Descriptive statements don't have this feature. You can fully understand and accept any descriptive statement without having any inclination to do anything. You can understand and concede the truth of the statement "if the button is pushed, many will die and those who don't will be miserable; if it isn't pushed, nobody will die and everybody will be happy" without having the slightest inclination to abstain from pushing the button. But you can't understand "you morally shouldn't push the button (since, e.g., it will kill innocent people)" without having such an inclination. So descriptive statements lack, whereas ethical statements have, the property that, if understood and believed true, they necessarily impel one towards certain acts.

Since correct ethical statements, supposing them to exist, necessarily incline those who understand and agree with them to behave in certain ways, and since descriptive statements *don't* have this feature, ethical

statements aren't descriptive statements. This means that ethical statements don't describe the world—that, in other words, they don't say how things are. This means that they don't say *anything*.

Analysis of Argument #5 Contrary to what this argument assumes, ethical statements *don't* necessarily incline those who agree with them to behave in certain ways. And, contrary to what this argument assumes, so far as ethical statements *do* incline people who agree with them to behave in certain ways, it is *only* to the extent that such statements are descriptive in nature.

Many people do things they know to be ethically wrong, and oftentimes haven't the slightest inclination not to do them. In many cases, it is the very fact that those acts are wrong that gives them their appeal.[375]

So far as I know, there is no empirical evidence that, whenever somebody deliberately violates an ethical norm, he did it *despite* the pull of some desire to comply with that norm. Certainly, the norm-violators themselves aren't always conscious of there being any such pull. Sometimes they are; sometimes they aren't.

I'd be very surprised if criminologists and psychologists had found such a pull to be operating in the unconscious of every norm-breaker who wasn't consciously aware of it. People have no wish to turn a blind eye to what is good in themselves or, therefore, to suppress their knowledge of their virtuous impulses. It is the ugly, evil impulses that they don't want to know about and that they have an incentive to conceal from themselves (and, of course, from others). People *want* to believe that they're motivated by noble impulses and aren't motivated by ignoble ones. Therefore, the former impulses are the least likely, and the latter are the most likely, to end up in the unconscious.

The wrong path is the easy path. Unethical acts are acts of giving in. A person's default mode is to act unethically. This doesn't mean that all people are all totally immoral or even that *any* of them are. To be immoral is not for one's *default*-mode to be immoral action. It is for one to allow that mode to be operationalized; and few people allow it to be *completely* operationalized, though many come close.

"Absolute power corrupts absolutely," as Lord Acton said. When people

don't have to act morally, they don't. Why not? Because they don't want to. So it's simply not psychologically accurate to say that no sooner

is an ethical principle understood than one is disinclined to violate it. Indeed, the idea that we're irresistibly drawn towards ethical behavior is so contrary to so many obvious realities that a fair amount of wishful thinking must underlie acceptance of it. The ethical is what you don't want to do. If you wanted to do it, it wouldn't be ethical anymore; it would just be another form of recreation, albeit one that, like friendship and certain other forms of intimacy, happened to conduce to the welfare of others.

Also, contrary to what Argument #5 assumes, it's only to the extent that one has beliefs as to the consequences of a given act that one feels compelled, morally or otherwise, to perform it. You feel compelled to give money to charity X only to the extent that you believe that, in so doing, you'd be alleviating hunger (or some such). Knowledge of moral truths compels action only to the extent that it coincides with knowledge of descriptive truths. So the extent that it's true to say that, if there are moral truths, knowledge of them compels action, it's only because moral truths coincide with descriptive truths.

> ***Argument #6*** Ethical statements are nothing more than instruments of propaganda. They are ways of tricking people into accepting existing distributions of power and wealth. In Hitler's Germany, obeying the Führer was "good" and disobeying him was "bad"; killing Jews was "right" and protecting them was "wrong." In Stalin's Russia, questioning the Communist Party was "the ultimate treachery," and blindly following its dictates was "the highest excellence." In capitalist countries, governmental regulation of private property is often described as "a gross violation of one's most fundamental human rights," even if such regulation is needed to provide basic services for many people. By the same token, there's "nothing wrong" with adding more millions to one's already large fortune, even if doing so involves impoverishing others, stripping them of opportunities, and generally diminishing their quality of life.
>
> These statements endorse values that support existing power structures. They trick the have-nots into helping the haves; and they do it by condemning acts that would help the have-nots and by praising acts that

would help the haves. In general, ethical codes manipulate the have-nots into hurting themselves and helping the haves. This is their function, and they discharge this function by representing injustice as justice. So ethical pronouncements are manipulative lies, cynical propaganda, not noble truths or indeed truths of any kind.

Analysis of Argument #6 Not everything that everyone describes as moral is moral. The Nazis *said* that obeying Hitler was the supreme ethical imperative. But they were wrong, and most of them knew it. So, while it's true that people's *claims* about morality are often bogus propaganda, it doesn't follow that *actual* morality is anything of the sort.

Also, one can *use* a legitimate ethical principle for some evil purpose. If a police officer bludgeons somebody to death for speeding, and then justifies it by saying that speeding must be punished to keep our highways safe, he's putting a good moral principle (people shouldn't speed or otherwise endanger the welfare of others) to evil use. Evil regimes (e.g., that of Stalin) go out of their way to find legitimate moral underpinnings for their nefarious policies. So given only that moral principles can be *used* to trick the have-nots into acquiescing to their condition, it doesn't follow that moral principles are *themselves* to be identified with such trickery.

Chapter 22

The Nature of Subjecthood and the Connection between Self-Interest and Morality

1.0 Practical vs. categorical incentives to act morally

There are often pragmatic incentives to act morally. In order to stay out of jail or to move up the corporate ladder, it may be necessary at certain junctures to act morally. But *setting aside* such pragmatic considerations, is it in one's interest to act morally? In other words, is it *categorically* in one's interest to do so? The purpose of this chapter is to answer that question. That answer is a very qualified "yes."

1.1 Why it's to one's advantage to be immoral

The very purpose of the laws we create is to incentivize moral conduct and to dis-incentivize immoral conduct. Murder, assault, rape, theft, extortion, and other immoral acts are severely punished. Becasue in its more extreme forms, non-cooperation with others is therefore punished by law, the pursuits that are legally available to one tend to involve behaviors that are, to some degree at least, conducive to the welfare of others. And, of course, even when it isn't required by law, it's often to one's practical advantage to treat others in an ethical manner.

But, quite obviously, it's also frequently to one's advantage to act immorally towards others. It's usually to one's advantage to curry favor with those in power. Those in power want you to serve their interests, and their interests are seldom those of morality.

1.2 Some intrinsic and extrinsic connections between happiness and immorality

No viable legal system could possibly dis-incentivize *all* immoral behavior. Laws are blunt instruments. And they *should* be blunt instruments, at least in

some ways. A legal system that eliminated all forms of immorality would, in so doing, penalize all forms of aggressive, anti-social, and self-interested behavior, except in the relatively few cases where it benefited others.

This means that, were all human conduct to fall within the scope of such a legal system, there would be no heroes, only zeros. There would be no Beethovens, no Einsteins, no Freuds (Sigmund or Anna), no Mendelssohns (Felix or Fanny), no Emily Dickinsons, no Madame Curies, and no Emily Brontes. There would be no geniuses, no conquerors—only gutless, faceless, brainless bureaucrats. This is because, more often than not, a certain amount of aggression and narcissism underlies great accomplishment. In a world of completely moral people, there would be no innovators, no creators—only milquetoast do-gooders and innocuous squares.

To be sure, in a world governed such a legal system, there would be no murder, theft, etc. But it wouldn't be a good world, since nobody would flourish in it. Nothing is more evil than preventing people from flourishing. Given that a *certain* degree of immorality is inherent in practically any great accomplishment, it follows that a legal system that made it completely impossible to act at all immorally would itself be supremely immoral.

Also, people living under the jurisdiction of such a legal system would have no freedom of choice. A world where it isn't even one's option to act immorally is a world where one doesn't have *any* options. Such a world obviously wouldn't be a good one. Therefore, a legal system that made all immorality impossible

would strip people of their autonomy, and, for this reason, would itself be immoral. Thus, there is at best a circumstantial connection between morality and self-interest: in some circumstances it is to one's advantage to act morally; in others it isn't.

But even this seemingly modest statement exaggerates the connection between morality and self-interest. So far as one's reason for behaving ethically is that it is to one's advantage to do so, one's behavior has no moral worth and, therefore, is not itself moral.

A story will make it clear why this is so. Brown is a sadist and serial killer. His sadism and wish to hurt others is surpassed only by his wish to save his own skin. Brown is drafted into the military during wartime. He is put to work as an interrogator, torturer, and executioner. He doesn't know

whether he's serving the interests of justice in discharging his professional responsibilities, and were he to find out that he was not, that would only enhance his love of his job.

As it happens, a successful job-performance on Brown's part will have very positive consequences not just for his fellow citizens but for the world in general. His country is a liberal, tolerant, peaceful democracy. It was invaded by a totalitarian, intolerant nation whose intention is to reduce the citizens of every country besides its own to a condition of slavery. For this reason, if Brown's country were to lose the war, humanity would revert to a condition of barbarism. And because of the particularly barbaric methods employed by the invading country, Brown's nation is itself forced to act barbarically in order to protect itself. Therefore, sadly, Brown's country, and people the world over, need Brown to do his job and do it well.

And he does. But Brown's excellent job performance is fueled by hate and ill-will, not by a love of justice. Brown wants to hurt people, and his job is the only outlet he has for this urge. Brown's love of his work would diminish were he to find out that it improved the human condition, and that love would become more intense were he to find out that it worsened it.

In doing his job, Brown isn't acting in a morally meritorious fashion, the reason being that his conduct isn't driven by a desire to do the right thing. Brown is acting morally only in the sense that his overt behavior coincides with that of somebody who is acting ethically. But Brown's despicable motives drain his behavior of all moral worth, and there is a sense in which his behavior, though behaviorally indistinguishable from moral behavior, is not itself at all moral. In general, unless the agent's intention is to act morally, his behavior is not moral, even if, from an external point of view, it is indistinguishable from moral conduct.

1.3 Why it is to one's advantage to be immoral

All publicly accepted systems of morality are to the effect that, with a few narrowly defined exceptions, *no* act that it is to one's advantage to perform is permissible if it involves hurting someone else.

That's why there are deeply moral people who are unhappy *for the very reason* that they are so moral. The "moral" person, almost by definition, grossly undervalues his own needs and aspirations, and the moral act is often

the one that is the most self-undermining and inimical to one's own welfare. Thus, many people would be much happier if they were less moral. And even though I'll argue that *some* morality, of a *certain* kind, is inherent in success and even subject hood, it's patently obvious that *too* much morality hobbles people. It turns them into weaklings and conformists who can't act without receiving the approval of others. Since that approval depends more on the needs of the would-be approval-giver than it does on the needs of the person seeking approval, the result is that too much morality-driven deference to others turns one into a cripple.

But there's a deeper, less circumstantial connection between hyper-morality and unhappiness. A precondition for being profoundly moral is having a punitive and restrictive conscience: a conscience that allows one to do, feel, or think as one wishes *only* under narrowly defined circumstances—a conscience that holds one's own psychological needs, and therefore one's very self, in the deepest contempt. The conscience of a deeply moral person, therefore, is one that embodies an attitude of hate, contempt, and aggression towards a facet of oneself. Thus, the essence of morality—or of hyper-morality, in any case—is a certain self-hate. Since nobody *not* weighed down by such a conscience can embody a life of pure morality, perfect morality isn't even compatible with happiness. A fortiori it isn't a necessary or a sufficient condition for it.

There are many deeply immoral people who are truly happy. Of course, some immoral people are wracked by self-loathing. They're appalled at who they are and what they do. They mask their self-hatred with cheap thrills and meaningless validations extorted from lickspittles. But some people who, in at least some respects, are quite evil are in fact happy. They don't simply feel good; they aren't experiencing the fake happiness that comes with self-deception and cheap pleasures. They are authentically, integratedly happy.

2.0 Why it's to one's advantage to be moral

Given these facts, the obvious judgment to make is that morality is *not* inherent in happiness. But such a judgment would be a shallow one. Being happy consists, at least in part, in judging that one's life conforms to some ideal that one has and, therefore, that it is in compliance with some *value.*

Happiness isn't so much a feeling as it is a conviction that how one is coincides with how, relative to *some* value system or other, one should be. Thus, the concept of happiness must be understood at least partly in valuative terms. It is a moral concept.

There is happiness only where there is an enduring psychological subject—a mere body is not enough. There is an enduring subject only where there is a certain degree of self-consciousness.[376] More specifically, for there to be a subject—what Harry Frankfurt refers to as a *person*—there must be an alignment of higher-order desires with lower-order desires.[377]

You have a flabby, pitiful physique. If you go for a jog right now, you'll trim down. If you eat a bucket of ice cream, you'll flab out. You give into lassitude and eat the ice cream. You're ashamed. The fact that you're ashamed means that, in your own eyes, you've done something wrong. You've fallen short of an ideal. You've failed to live up to your own *values,* and you are thus unhappy. (In any case, you're unhappy vis-à-vis this particular situation, though perhaps not in general.)

To move forward, let's modify the story just told. As before, you want a trim physique and are confronted with the same choice as before. But this time, you go for the jog. You are proud of yourself, and you're also delighted at the fact that, thanks to your self-control, you are that much closer to your objective. You are happy (vis-à-vis this situation).

Thus, it is only to the extent that a creature *values* things that it makes sense to describe it as happy or unhappy. Anything that values anything *ipso facto* has a value system and thus a morality. And to be happy, a creature must not only have such a value system, but comply with it.

This doesn't mean that, in order to be happy, one must comply with *all* the value systems that one countenances. Any non-sociopathic human being sincerely countenances many values that he couldn't comply with completely without obliterating any chance of being happy. But a given creature's condition is one of happiness, as opposed to mere pleasantness, *only* if it's living up to certain values that it has; and such a creature's condition is one of unhappiness, as opposed to mere unpleasantness, *only* if it is falling short of certain values that it has.

3.0 Necessary versus sufficient conditions for happiness

In order to be happy, is it enough that one live up to one's values?

Yes. But it's easy to see why one might initially think otherwise. Happiness is partly based on what one *does* and, therefore, on things that are within one's control. But it's also based on things that aren't always in one's control (e.g., one's physical health). Being happy involves *behaving* in ways that one approves of—it involves *doing* the things that one believes it important to do. But it also involves *being* the way one wants to be, and it isn't always up to one to be, or not to be, in the requisite ways.

Smith is a great novelist who has fully capitalized on his talent. But, through no fault of his own, he has a terrible, chronic illness, which causes him constant pain. Is Smith happy? No. Is that because our analysis of happiness is wrong? Is that because we were wrong to identify happiness with a condition in one's actual self is in alignment with one's idealized self? No. It's because that alignment depends, not *just* on one's actions, but also on things that one may not be able to control (e.g., physical health).

So, while it is a fact that one must feel good to be happy, this is perfectly consistent with our analysis. According to that analysis, being happy is identical with, or at least involves, having the conviction that one is as one wants oneself to be. Smith wants his body to be healthy—but his body isn't healthy. He wants to

be free of pain—but he's wracked with pain.[378] So he falls short of his own benchmarks. It's irrelevant that it isn't his fault.

3.1 Happiness as an underlying structure that must be distinguished from the positive *feelings* to which it gives rise

One can be happy in some respects and unhappy in others. Smith is justly proud of his work and he finds his work exquisitely pleasurable. His constant physical pain obviously diminishes the quality of his life. But does it drain *all* happiness from it?

It does not follow. It's also possible that Smith is happy vis-à-vis his work but not vis-à-vis his health situation. "But that's preposterous," it will be said. "Happiness is a *feeling.* You're either happy or you aren't. And Smith isn't happy. That's why he's wailing all the time."

Some mental entities are structures. Others are events or states that occur within the limits set by such structures but, unlike them, are fundamentally ephemeral. Beliefs and values are structures. So are personality traits.

These things can change, of course. One's beliefs, even one's personality, can change. But they're not happenings. Joy and pleasure are in a different category. They're passing states. Unlike one's knowledge that 1 + 1 = 2, they're not structures, not even ephemeral ones. They're structure-internal events.

Happiness is a structure, not a feeling. In this respect it's more like one's knowledge that 1 + 1 = 2 than it is like a passing feeling of joy. Mental structures are expressed by mental happenings and also by overt behaviors. But one's knowledge that 1 + 1 = 2 can't be identified with any of the events or passing conditions to which it gives rise. And while happiness is likely to give rise to joy and pleasure, it can no more be identified with them than one's belief that 1 + 1 = 2 can be identified with one's recognition that Bill has two cars (given your just-acquired knowledge that he has one car in Idaho, a second car in Montana, and no other cars). Just as that recognition is an effect of, and is therefore distinct from, that knowledge, so joy and pleasure are effects of happiness, and are therefore distinct from it.

Given that happiness is a structure, it no longer seems unreasonable to say *either* that one can be happy in some respects while being unhappy in others *or* that happiness is a *judgment.* (In this context, by "judgment," I don't mean the *act* of judgment. I mean that enduring conviction that one has as a result of a judgment in whose correctness one firmly believes.) Judgments, and therefore value-judgments, are not feelings. They are mental structures, not mental events or transient mental conditions, even though they obviously *result* in such events and conditions.

Therefore, value-judgments can result in joy, sadness, pleasure, and displeasure. And a positive value-judgment about oneself would, for obvious reasons, be likely to result in pleasant feelings, and a negative one would be likely to result in unpleasant feelings. Thus, the thesis that (un)happiness is, or at least consists in part of, a value-judgment about oneself is consistent with the tight connection between happiness, on the one hand, and feelings of (dis)pleasure, etc., on the other.

Moreover, that thesis is consistent with the supposition that one can be happy in some respects (e.g., about one's literary accomplishments) and unhappy in others (e.g., one's health or family situation). For one can obviously judge that one's life meets or exceeds certain standards, while falling short of others.

3.2 Subjecthood a precondition for happiness; self-consciousness a precondition for subjecthood; norm-drivenness a precondition for self-consciousness

We have values, and in many cases our actions are driven by the values that we have. There is a gap between what we want, and therefore value, and what there is. Wishing to narrow the gap, we act. If an act is driven by a wish to make reality conform in some respect to a value, let us say that is "value-driven." We'll use the same term to describe the thoughts out of which value-driven acts arise. In this section, we'll see that it is inherent

in the concept of action that it be value-driven, and we'll also see that value-drivenness is of the essence of subjecthood.

Rocks can't be happy. Why not? For starters, they aren't sentient.

Sentience is necessary for happiness. But it isn't sufficient. Not all animate beings can be happy. Only animate beings meeting certain conditions have that privilege. Not all animate beings are *subjects.* The difference between animate beings that are subjects and those that are not is that the former, unlike the latter, are able to regulate their own lives according to *values* that they have. As we'll see, a certain degree of self-consciousness is needed to meet this condition.

We are inclined to think that non-human animals (e.g., cats, dogs, hamsters) are without values. This isn't true. There are certain kinds of values that they lack that we have. But it is inherent in the very nature of subjecthood that its activity be value-driven. Cats, dogs, etc., are indeed subjects. (Not all animate creatures are subjects.)

We're also inclined to think that they are completely lacking in self-consciousness. This isn't true. Echoing what we just said, there is an important kind of self-consciousness that we have that they lack. But in order for a series of mental states to have the cohesiveness needed to constitute a single mind, as opposed to a mere sequence of mental states, it is necessary that some of them embody some awareness of some of the others.

Let X be a creature of one-thousandth your intelligence that is constantly happy. X's life consists of one ecstasy after another. But none of these ecstasies embodies any intelligence, any connectedness to other life-forms, any connectedness to any ideals or values. X's life is a series of one-dimensional pleasures after another.

Further, let us suppose, these experiences don't build on one another. X feels good at time t_1; it feels good again at time t_2; and again at time t_3; and so on. It may be that X feels good in different ways on all of these different occasions. Maybe at t_1 it feels good in the way that somebody eating a delicious chocolate bar feels good. Maybe at time t_2, it feels good in the way that stepping into a hot bath feels good. And so on. But it doesn't synthesize these various experiences; it does not, on the basis of them, define objectives for itself that it then attempts to realize. No—everything is handed to it. Life is just one good feeling after another. It doesn't have to aspire towards anything. Nor does it want to. Nor indeed does it want *anything.* X just *feels.* And it feels good. But not just good—ecstatic.

But is it *happy?* X isn't *un*happy. But it isn't happy either. The concept of happiness is inapplicable. Rocks are neither happy nor unhappy. They don't meet the prerequisites for happiness or unhappiness. One such prerequisite is being animate. Rocks don't satisfy that condition. X does. But there are other such prerequisites that X fails to fulfill.

What's missing? First of all, it isn't clear what "X" refers to. "Joe Biden" refers to a person. That person has a body, and that body is integrally involved in that person's existence. But "Joe Biden" refers to the *person* whose existence is maintained, in some way or other, by that body, not to that body per se. If that body had been born without a brain, surely Joe Biden would never have existed (or so I argue in Chapter 16). And if that body slips into a vegetative coma, and it's known beyond any doubt that there is no possibility of any resumption of mental functioning, Joe Biden is "gone," in some very significant sense. But the body isn't gone. So "Joe Biden" does not refer to a body, but to a person—an animate entity—that depends on a certain body. ("Joe Biden" refers to an embodied person. But it doesn't refer to a body. Joe Biden's body will outlast Joe Biden, sadly.)

We give names to pets. And in most cases these names typically function like "Joe Biden." "Fido" doesn't refer to a body; it refers to an embodied presence—a friendly canine "personality" that, like my personality and yours, depends for its existence in some important way on that of a certain body.

Oftentimes, we use proper names to refer to purely physical entities. We give names to mountains, valleys, bays, rocks, trees, geysers, waterfalls, etc. Such names aren't referring to things that, like you and me, are embodied; such names refer to the bodies themselves, not to the things that have them. So "Mt. Everest" refers, not to something embodied in a certain wrinkle of the Earth's surface, but to that wrinkle itself; and that name thus seems to be functioning in a very different way from "Joe Biden" and "Fido."

How is "X" functioning? It's obvious that some *body* is associated with X in some way; and that body could obviously be given a name. But is there something that *has* that body, in the way that you have your body and I have mine?

How is "X" functioning? It's obvious that some *body* is associated with X in some way. And, of course, that body could be given a name. But is there something that *has* that body, in the way that you have your body and I have

mine?

We must take care not to impute our mental states to X. When *we* feel pleasure, we don't *just* feel pleasure. We put that feeling into a larger context. We think about where it came from; what it's leading to; what it accompanies, how it bears on our objectives; whether we should feel guilty about it; whether it's a healthy or an unhealthy development; whether other people can tell that we're feeling it; whether, supposing it pleasurable, we'll be able to recreate it; whether we'll be able to recreate it without jeopardizing the welfare of people we care about (or who don't care about but feel we have a moral obligation towards); etc.

When *we* have an experience, even one we consider base or trivial, we look it at through the lens of a preexisting structure, formed partly on the basis of other experiences, and that experience may, in its turn, become a part of such a structure. If it's exceptionally (un)pleasant, we'll try to remember what caused it so as to (not) recreate it.

Our mental lives are therefore cumulative; and inherent in this cumulativeness is a certain self-consciousness. It is sometimes said that only human beings are self-conscious. If this is true, it is only in a narrow sense of the word "self-conscious." The mental life of a dog or a cat is obviously cumulative in much the same way as ours. Because dogs and cats aren't as good at we are at thinking abstractly, what they know is much more tightly tethered to their perceptual experiences than what we know is tethered to ours. But their psychological lives don't consist of mutually disconnected, isolated mental episodes. Like us, they learn, develop attitudes, and have desires. And, possibly, they even have objectives.

But none of this holds of X. In X's case, one state is followed by another; and that one is followed by a third; and so on indefinitely. But X's mental life isn't cumulative, at least not in the sense just described. True—X is animate. In any case, the body associated with the expression "X" is animated by certain mental states. But it isn't so clear that there is a subject *in* that body. In fact, there seems *not* to be.

For there to be a subject, it isn't enough that there be mental events. There must be a structure that not only responds to these events but does so in accordance with views as to how it *should* respond to them. Mere sequences of mental events are not enough for subject-hood. *For a subject to exist,*

mental events must be responded to in a norm-driven manner. (This claim is crucial and will be clarified and defended in a moment.) And it is only where there is self-consciousness that norm-driven responses are possible. It follows that norm-drivenness is constitutive of subjecthood.

3.2.1 A defense of the italicized claim

It's one thing for your *belief* that Max has four cars to be the cause of your belief that Max has more than two cars. It's a very different thing for your *believing* that Max has four cars to lead to your believing that he has more than two cars.

A story will make this clear. At time t, you learn that your friend Max has exactly four cars. Let M_1 be this belief. At time t*, you come to believe that Max has more than two cars. Let M_2 be the mental event involved in your learning *this* fact.

M_1 and M_2 are presumably identical with, or constituted by some brain-structures of some kind. Let B_1 and B_2 be these brain-structures.

B_1 has many properties. It has the property of mediating a belief of a certain kind. But it also has various non-psychological properties. Some of these properties are such that, if B_1 didn't have them, it would for that very reason be incapable of mediating a mental entity of any kind, let alone a belief that Max has four cars. But others are such that, were B_1 to lack them, it would not, at least not for that reason alone, undergo any psychological change at all. Let n be B_1's exact temperature, and let m be some temperature that is only one millionth of a degree greater than n. Were B_1's temperature to shift from n to m, it would not, in virtue of that fact, change in respect of its psychological characteristics. Thus, B_1's having some *exact* temperature is irrelevant to its being a belief that Max has four cars, and so is B_1's having some *specific* shape or *specific* microstructure. The same thing is true of B_2 and, indeed, of any physical entity that mediates some mental event or condition.

During what follows, we'll suppose that B_1 has shape X and that it's

having that exact shape has nothing to do with its mediating a belief that Max four cars.

At time t*, because of some strange quirk on your part, B_1's having shape X—*not* its being a belief that Max has four cars—brings about M_2. (The exact nature of the causal connection doesn't matter. But, if it helps, we may suppose that, because of its shape, B_1 sets off a migraine which, because it is so intense, leads to your having hallucinations of Max showing you titles to three cars of his.)

B_1 brought about B_2. B_1 and B_2 are, respectively, your belief that Max has four cars and your belief that he has more than two cars. So your belief that Max has four cars brought about your belief that he has more than two cars. But B_1's *being* a belief that Max has four cars had nothing to do with B_2 or *a fortiori* with B_2's being a belief of any sort. So your believing that Max has four cars has nothing to do with your believing that he has more than two cars.

A second story will clarify the first. I learn that Max has four cars. (Let M_3 be my belief that Max has four cars.) I am jealous. Out of jealousy, I try to throw a rock at one Max's cars. I miss. The rock hits Smith in the head. (Smith is some person who happens to know Max.) The resulting head-trauma leads Smith to believe that Max has more than two cars. The connection between M_1 and M_2 is similar to the connection between my belief that Max has four cars and your consequent belief that he has more than two cars.

Of course, M_1 and M_2 belong to the same mind, whereas M_3 and M_4 don't. But it isn't *because* M_1 brings M_2 about that they belong to the same mind. M_3 brings M_4 about, but they don't belong to the same mind. So why do M_1 and M_2 belong to the same mind?

"M_1 and M_2 are identical with events that occur in the same *brain.* And that's why they belong to the same mind."

That's not a very good answer. A given brain could, theoretically at least, house different minds. It may be that, in actuality, no brain houses two

different minds. But it's clearly an empirical question. There's nothing conceptually wrong with the idea. In fact, there is some experimental evidence that, at least arguably, suggests that different minds *can* co-occupy some one brain.[379] That same evidence suggests that, were it not for the existence of quite fragile neural structures, the subject housed by a given brain would powder into a multiplicity of distinct subjects.

Second, the reason that we see brains as constitutive of minds is that we know on independent grounds that the mental events mediated by a given brain tend to constitute some one mind. In the (1987) movie *Like Father, Like Son,* a stodgy father (played by Dudley Moore) switches bodies with his fun-loving son (played by Kurt Cameron). Events of this kind may well be impossible; but they aren't impossible in the same way that square circles are impossible. That is to say, they are not *logically* impossible. I'm pretty sure that I won't wake up in somebody else's body tomorrow. But that belief of mine isn't in the same category as my belief that there are no square circles. The former, unlike the latter, is an empirical belief. I cannot coherently imagine having had life-experiences that would warrant my believing that there are square circles. I *can* coherently imagine having had life-experiences that would warrant my believing that I'd wake up in a body different from the one I now occupy. And were events of this kind to happen, we wouldn't see brains as individuative of minds. This means that brains are individuative of minds only in a parasitic sense; which, in its turn, means that, ultimately, brains aren't individuative of minds at all.

"Maybe the reason that M_1 and M_2 belong to the same mind is that they belong to the same body." This answer fails for the same reason as the previous one.

"Maybe the reason that M_1 and M_2 belong to the same mind is that M_1 and M_2 belong to the same soul."

This answer is vacant, since for two events to "belong to the same soul" is, tautologously, for them to belong to the same mind.

"Maybe the reason that M_1 and M_2 belong to the same mind is that M_1 causes M_2 to occur."

But M_3 causes M_4 to occur, and M_3 and M_4 belong to different minds. Also, within a single mind, there can be simultaneous, and therefore causally unconnected events. And, also not everything that went in my mind at age three is causally connected to everything that goes in it now.

Nonetheless, this last answer is on the right track, as another story will make clear. I come to believe that Max has four cars. (Let M_5 be this event, and let B_5 be the brain-entity with which it is identical.) I validly infer that Max has more than two cars. (Let M_6 be this belief and let B_6 be the brain-entity identical therewith.) In this case, my belief that Max has four cars brings about my belief that Max has more than two cars, and—what is different, as we have seen—my *believing* that Max has four cars brings about my believing that Max has more than two cars.

But there's more to it than that. It wasn't in some haphazard or unprincipled way that my believing the one thing led to my believing the other. My believing that Max has four cars could have led to a jealousyinduced psychotic break during which I had hallucinations of Max showing me the titles to three cars of his. For argument's sake, let's suppose that, on the basis of those hallucinations, I believed that Max had more than two cars. Under those circumstances, my (pre-hallucination) belief that Max had four cars would indeed have led to my belief that Max has more than two cars. But the transition from the one belief to the other wouldn't have been driven.

But that isn't what happened. In actuality I knew that Max had four cars and, driven by an understanding of the implications of the fact, I rightly inferred that Max must have more than two cars. That transition was *norm*-driven. I recognized that Max must have more than two cars, given that he has four; it was by way of this recognition that M_1 brought about M_2, *this* being why that transition was norm-driven.

People make invalid inferences of course. But those are norm-driven. (They're driven by the wrong norms.) I believe that all dogs are mammals and invalidly infer from this that all mammals are dogs. This inference, though invalid, is norm-driven. The norm just happens to be a bad one.

There are norm-driven interactions only where there is self-

consciousness. Consider the inference leading from M_5 to M_6. An awareness of M_5's content was involved, and so was an awareness of what, given that content, ought to be believed. M_6's being a norm-driven response to M_5 thus involved one mental event of mine embodying an awareness of the content of some other mental event of mine.

My eating cow's brain and finding it repellant leads to my avoiding eating cow brain and also to my avoiding things that, I believe, are likely to involve my eating cow brain (e.g., it leads to my declining dinner-invitations that I'd otherwise accept).

These developments are driven by a knowledge on my part of norms. I know that I don't want a repeat of that horrible experience. I infer in a reasonable, and therefore norm-driven, manner that, if I eat cow-brain again, I will re-experience the horror. My fear of eating cow-brain embodies that legitimate inference. My making that inference involves forming a belief as to what I should do *given* what it is that I want (viz. to not undergo a repeat of that horrible experience) and *given* what it is that I believe (viz. that, were I to eat cowbrain again, I'd probably undergo a repeat of that experience). Thus, that inference, and the reaction at whose center it lies, involves a certain self-consciousness.

Whenever one mental state is *based on* another, the first embodies a content-driven response to the second, and it also, for that very reason, embodies a norm-driven response to it. I hate eating cow-brains. I believe that I'll be forced to eat cow-brains if I get within a ten-mile radius of Sally's house. I become nervous when the car I'm in, which Sally's friend Mary is driving, suddenly and unexpectedly veers in the direction of Sally's house. My nervousness is *based on* my belief that I'll have to eat cow-brains if I get too close to Sally's house. (Part of what this means, though not all of it, is that, if B_7 is that belief of mine and B_8 is my feeling of nervousness, it isn't B_7's having this or that shape or mass or color that brings about, and is instead B_7's having a certain content that does so.) To say that it's "based on it" is to say that it embodies a view—not necessarily a good one—as to how one *ought* to react to the (mis)information embodied in it. So for B_8 to be "based on" B_7 is for the former to be, or embody, the outcome of a norm-

driven response to the latter. In general, content-driven reactions are norm-driven reactions.

Given a set of mental events, none of which was a norm-driven response to any of the others, those events would not collectively form a single mind or subject. There are no content-driven reactions where there are no norm-driven reactions. Where there are no content-driven reactions, nothing has anything to do with anything else. Let $m_1...m_n$ be a set of mental events having contents $c_1...c_n$. (So if m_1 is a belief that grass is green, then m_1's content is the proposition that grass is green.) And let $b_1...b_n$ the corresponding brain-entities. Even if b_1 is what causes b_2 to occur, and b_2 is what causes b_3 t occur, and so on, so long as b_{37}'s *being a belief that arsenic is poisonous* isn't related to b_{86}'s *being an intention to avoid drinking the contents of the bottle labeled WARNING: CONTAINS ARSENIC*—so long as, in general, bi's having content ci has nothing to do with bk's having content ck—then, even though those brain-states are causally (inter)connected, the corresponding mental states have nothing to do with one another. Thus, where there are no content-driven reactions, no mental state has anything to do with any other. Mental states cohere, and form a single subject, only if some of them are content-driven responses to others. This means that there are subjects only where there are content-driven and therefore norm-driven reactions. So norm-drivenness is the essence of subjecthood. The essence of subjecthood is thus a concern for what *ought* to be the case.

"But the 'ought' in question surely isn't a *moral* ought."

Why not?

"Because legitimate inferences—i.e., inferences that embody due regard for the appropriate norms—can be made for purely self-interested reasons. One can use one's mathematical skill to steal money, etc."

There's no denying that profoundly evil people can be very intelligent. But we'll see that norm-drivenness in the *moral* sense is to be understood in terms of narrower form of norm-drivenness that we've seen to be constitutive of subjecthood.

One last point: Isn't my aversion to eating cow brains purely an

emotional or instinctual reacting to my unpleasant dining experience? Yes—but only because instinctual and emotional reactions are norm-driven. The difference between a "cool, logical reaction" and "a subjective, emotion-driven reaction" isn't that the latter isn't norm-driven, whereas the former is; it's that the latter is driven by a norm that, were one not blinded by passion, one would question. If, driven by fear and hate, Smith decides that Larry must be dangerous, the reason being that Larry has the same hair color as Larry's evil brother-in-law, Larry's emotion-driven reaction is norm-driven. The problem is that the norm isn't a good one. A reaction that isn't driven by *any* norm isn't prejudiced, the reason being that prejudice *is* acceptance of illegitimate norms. If a reaction is emotional and otherwise 'irrational', it is for that very reason, norm-driven. (I put "irrational" in scare-quotes since I don't believe that emotional reactions are necessarily irrational ones. They *can* be irrational, of course.[380] But emotions can give one conscious access to insights that would otherwise be submerged in one's unconscious. See Chapter 20.)

3.3 Why an excessively narrow conception of self-consciousness obscures the nature of subjecthood and morality

Animals are obviously capable of self-consciousness of the kind just described.[381] Animals don't simply *react* to their experiences; their reactions are hewed to the contents of those experiences and embody the same sort of sensitivity to the demands of normativity embodied in our reactions to our experiences. Of course, we are in some respects more intelligent than animals. We're obviously better at thinking abstractly than they are. But an animal's experience has the relevant sort of cumulativeness and, therefore, embodies the relevant sort of self-consciousness, which is why animals clearly are subjects.

Most non-human animals, maybe all of them, clearly lack a *certain* kind of self-consciousness that we have. But they don't lack the kind needed for subjecthood. But X *does* lack just that. And that's why X isn't really a subject. There's a body, and there are mental states. But there's no subject

that occupies that body that "has" those mental states. *Nothing* has them. They happen *in* a body, but they don't happen *to,* or belong *to,* anything. They just happen.

X isn't a subject. You are. What's the difference? X's mental states merely succeed one another. There are, or may be, causal relations between them; X's having a certain feeling at t_1 may cause it to have a certain feeling at t_2. But these causal relations aren't normative in nature; given a particular mental state M had by X, the nature of the mental states following it isn't a function of M's *content.* Suppose that X is in mental states $M_1...M_n$ and times $t_1...t_n$, respectively. M_1 may *cause* M_2 to occur, and M_2 may *cause* M_3 to occur, and so on. But that isn't enough for these states to be constitutive of some one subject. Unless some awareness of what, givenM_1's *content,* is *appropriate* is involved in M_1's bringing M_2 into existence, the sequence consisting of those two events is *just* a sequence; those two events don't form some one mind. They're just two distinct mental states that are associated with some body. For a multiplicity of mental states associated with some one body to constitute a single mind, it is necessary that an awareness of what is *appropriate* given the *contents* of existing mental states—of, in other words, what is *demanded* by the contents of those states—be what drives the sequence of events associated with this body. Thus, it is necessary not only that those states be causally, but also *normatively,* cohesive. Their mutual causal relations must embody a sensitivity to norms and, therefore, to value. Value-drivenness is the glue that makes a multiplicity of otherwise disconnected mental states constitute a single subject. So wherever there is a *subject*—as opposed to a sequence of mental states that, although associated with some body, don't jointly form a single mind—there is sensitivity to value. The essence of subjecthood is sensitivity to value and, therefore, to a kind of proto-morality, if not a *bona fide* morality. Morality thus has much deeper roots than we are inclined to think. When it is said that only human beings have values, that is either false or what is being said is that only human beings have values of some very specific kind. For there to be a subject, it isn't enough that a multiplicity of different mental states be associated with some one body. It is necessary that those states cohere, not just causally, but normatively. Thus, a sensitivity to value, and thus a kind of

morality, lies at the root of *all* subjecthood.

3.4 Morality = respect for subjects qua subjects

"But not all values are moral values," it will be said. "Ted Bundy valued killing people. But that wasn't a moral value." This is false. Morality is about respect for subjects. Moral values are those that must be complied with if subjects are to be allowed to exist—if, in other words, multiplicities of mental states are to have the integrity needed for them to constitute single minds. Moral values are therefore to be understood in terms of the values embedded in *any* such multiplicity of mental states. Let us turn to this right now.

3.5 Flourishing as the one thing that has non-derivative moral value

Let's start with a story. Smith would very much like to write novels. He's extremely talented. He'd be happy if he spent his days writing. But Smith feels that, if he were to do so, he'd be acting immorally. "There are so many people who are suffering," he tells himself, "so many people who don't have food, water, plumbing, etc. So how can I justify writing novels all day, when others are deprived in such fundamental ways? Wouldn't my doing so be nothing more than callous self-indulgence?"

Driven by his admirable concern for the dictates of morality, Smith spends his days installing sewage systems in poor countries, giving food to the hungry, etc. He doesn't find this stimulating or enjoyable. He hates it in fact, and his life is a living hell. He does it because he believes (rightly, no doubt) that, in doing it, he is acting morally.

Smith's situation isn't representative of that of all altruists. Most people who dedicate their lives to the welfare of others find it fulfilling; for them, helping others is what novel writing is (or would be, were he to let himself do it) for Smith. But Smith isn't one of these people. Smith's life doesn't "work for him," as some self-help gurus say. He knows intellectually that what he's doing is morally good. But from the viewpoint that is embodied in his *emotions,* as opposed to the viewpoint embodied in his intellectual *judgments* as to what is moral, his life is a waking death. And *his* viewpoint is that of his

heart. His intellect embodies a viewpoint that, despite his best efforts, he could never make his own.

Let's contrast Smith's situation with Brown's. Like Smith, Brown is a talented novelist. But unlike Smith, Brown permits himself to write novels. As a result, he is happy; his life "works" for him. Unlike Smith, Brown is a self-interested scoundrel. He doesn't care about others. This isn't because he *can't* care about others. Every now and then he develops feelings for someone and cares about that person. So, although he's a self-involved narcissist, he isn't so emotionally stunted that he can't have the degree of empathy requisite to have meaningful and fulfilling relations with others. In fact, his social and romantic life is very fulfilling. But when it comes to people who fall outside the sphere of his personal relations, his attitude is a cold and calculating one. Where such people are concerned, he does what he has to do and doesn't worry about the consequences.

Because Brown has allowed himself to pursue his dream of novel writing, and because, as a result, he's allowed himself to develop his talent for it, he wakes up every morning with a smile. He's happy to greet the day. He looks forward to another day's work.

Although Brown is a selfish swine, he *does* have artistic integrity. He wants his work to be as good as it can be. To this end, he listens very carefully to what even his most hostile critics have to say and makes a serious effort to incorporate their points into his work (at least when those points have merit and aren't mere vituperations). So Brown is legitimate *at what he does.* It may be that he isn't *morally* legitimate. But, as a novelist, he isn't a fraud; he's the real deal.

Whose life is better, Smith's or Brown's? It's obvious that Smith's life is the *morally* better one. But what I'm asking is: whose life is actually a richer, better, less empty one? The answer is clear: Brown's. It would be nice if Smith's life were the better one. But it isn't. Brown has a good life. Smith does not.

Smith's life is *morally* good, of course. But it isn't good in the relevant way—in the way that is relevant to his welfare, his flourishing. In being morally good, Smith's life is conforming to a standard that is totally external to itself—that has no basis in its own logic. Smith's life is a bad one, and not

just bad in a narrowly hedonistic sense. A certain degree of asceticism and self-discipline is needed to accomplish anything of value. But there is a difference between self-discipline and self-torture. Brown is self-disciplined. He works hard to hone his skills as a novelist. It is this self-discipline that prevents Brown from spending his days consuming mind-bending narcotics and indulging in other hedonistic, debauched pleasures that would weaken his artistic ability and thus reduce him as a human being.

Like Brown, Smith is self-disciplined. But he's self-disciplined in a different way. Brown's self-discipline is about making the sacrifices needed to be a great novelist and, thus, to flourish. Smith's self-discipline is about *not* doing what he has to do to flourish; it's about denying who he is, not enhancing it. And while Smith is obviously a person of principle and must be admired for that, his life is not a good one.

3.6 The good life—one in which one is true to one's self, not to external conceptions of morality

Some ethicists have tried to show that, ultimately, lives such as Brown's are hollow and that lives such as Smith's are meaningful and good. But the facts don't bear this out. Having a good life is about having a life that is true to who you are. Smith's life is about being untrue to who he is; the conceit underlying his every action is that his own desires and, therefore, the very self they spring from, are an abomination. No one who sees who he is as an abomination and spends his life penalizing himself has a good life, no matter how much joy his self-flagellations bring to others.

It isn't ethical to demand of people that they suppress their true selves. An ethical system that does so is an evil one. Since it is the purpose of an ethical system to promote good, not evil, any such ethical system is an incoherent one. And yet most ethical systems are in fact evil and incoherent in exactly this way. The subject's own aspirations, though sometimes *tolerated,* are seen as having no *moral* value; all that counts form a moral perspective is what one does for others.

"But what if people's true selves are evil? In *that* case, surely, an ethical system that demands a suppression of self needn't be an incoherent one."

Supposing (i) that we're evil and (ii) that this fact about us is inherent in who we really are, then (iii) a viable ethical system mustn't attempt to excise it from us, and must instead find a relatively harmless way of redirecting it. An ethical system must be rejected if one must mutilate oneself to comply with it.

It should also be kept in mind that, if a given ethical system mutilates those who try to comply with it, those who advocate that system are probably well aware of that fact, and their advocacy of that system is almost assuredly rooted in their wish to hurt and diminish other human beings. (Moral censure is one of the few outlets for aggression permitted us by civilized existence.) It's a distinct possibility that, given what human beings are, *no* completely coherent ethical system is a viable one. It may be that, if people are to be prevented from destroying one another, they must be restrained by ethical systems that diminish them—by ethical systems that are themselves unethical and, therefore, incoherent. But, speaking from the viewpoint of logic, as opposed to practicality, no ethical system that requires people to obliterate themselves, even partially, can be regarded as a coherent one. Ultimately, the *only* thing of value is flourishing. A life, such as Smith's, that is "moral" is, sadly, a worthless one. A life of self-denial—not of appropriate self-discipline, but of self-effacement—is, for that very reason, constructed around the premise that it itself is worthless and, for that reason, aspires to validate that premise through continual self-denigration.

3.7 Flourishing ≠ self-indulgence

No moral system can be viable if it fails to distinguish flourishing, which obviously has moral value, from mere self-indulgence, which does not. No moral system can be viable if the central premise of it is that, to have a life worth living, one must negate who one is. The central tenet of any viable moral theory is that, ultimately, the one and only thing that is of value is flourishing.

In many cases, flourishing involves helping others. Most people, in order to have good lives, have to be connected to others in cooperative and supportive ways. And for many people, flourishing means spending one's days in the service of others. But that isn't what it means to everyone. And if it isn't what it means to you, your dedicating your life to an ethic of altruism,

though laudable, will strip your life of any *intrinsic* value. Your life will still have *value;* it will still have *moral* value. But its value will lie in the fact that it makes it easier for others to flourish. Any value that it has will lies in the benefits it yields. It itself will be a horror.

What is generally described as "morality" is *a* kind of morality—a very important kind. But, contrary to what is generally alleged, it is actually a derivative form of morality. Non-derivative moral value lies in flourishing—nothing else. It is only a derivative, instrumental sort of moral value that lies in altruism (except in so far as people flourish by being altruistic—which most do to *some* extent, and some to a very great extent). We will now develop this idea.

4.0 The morality of duty vs. the morality of aspiration

Some ethicists distinguish between what they call "the morality of duty" and "the morality of aspiration." The morality of duty is what would ordinarily be described as morality. It consists of the directives that forbid us, except under special circumstances, from killing, stealing, lying, etc.

The so-called morality of aspiration is a very different sort of thing. It consists of the directives that one must follow if one is to have a good life—a life that is good in the sense in which *Brown's* life is good, not in the alienated sense in which *Smith's* life is good.

Two questions arise. First, is it right to speak of the *morality* of aspiration? Given how different the so-called "morality" of aspiration is from the acts of self-sacrifice that we would customarily see as paragons of morality—nay, given how antithetical the two are—it is by no means obvious that, in the expression "the morality of aspiration," the word "morality" shouldn't be replaced with some other.

Second, what is the connection between these two sorts of morality (or, if the morality of aspiration isn't really a morality at all, between these two things)? Do the two sorts of morality ever coalesce? Is compliance with the morality of duty *necessary* for flourishing (that is, for compliance with the morality of aspiration)? Or, on the contrary, are they incompatible? Or are

any connections between them, whether positive or negative, purely circumstantial?

We've already answered the first question—the answer to it is "yes." We saw this when discussing Smith and Brown. Smith's life is one of constant torment. He is a person of talent and goodness who is rewarded for his troubles with unending distress. So far as his life has any value, it lies in the fact that it brings joy to others. Thus, given what Smith has chosen to do with his life, his life per se is worthless, its only value being that it conduces to the welfare of others. Smith himself admits this—in his actions, if not his words. In persistently disregarding his own needs, he's saying: "I live but for others; my own life is but a means to their ends." Which is a way of saying: "my life has no intrinsic value and, therefore, no *real* value."

The only *intrinsically* valuable thing is flourishing. (It follows that what is intrinsically valuable for me is different from what is intrinsically valuable for you. My flourishing is what is intrinsically valuable for me; your flourishing is what is intrinsically valuable for you.) Acting morally at least sometimes consists of acting in a way that's conducive to the flourishing of others. This is a given. It would therefore make no sense to say that flourishing itself has no value. In other words, the following statement makes no sense[382]:

(MC) It is morally valuable to help others flourishing; but a person's flourishing does not itself have moral worth.

If helping others flourish has moral value, which it obviously does, that moral value must be derivative of the moral value had by flourishing itself. MC is therefore comparable to the statement:

(MC*) Effective medicines have value, the reason being that it enables people to have good health. But good health does not *itself* value.

MC* is nonsense; and so, therefore, is MC.

4.1 The connection between the two kinds of morality

But we have to yet answer the question: "what is the connection between the

morality of duty and the morality of aspiration?" Let us turn to this now.

Acts that we typically describe as "immoral" are those that are incompatible with the interests of others. "Immoral" acts disrupt a social equilibrium on which people depend. Acts that we typically describe as "moral" are those that don't violate the interests of others. But "immorality" can't be defined as "the thwarting of other people's interests"; there's more to it than that, and the same thing *mutatis mutandis* is true of the term "morality."

Smith is a person. He is a valued member of the community. I personally like him. But I kill him. Why? Because I want his money so that I can buy a third yacht. My killing Smith is "immoral." But why? Is it because I violated the interests of another human being? That's surely a very important part of the answer. But it isn't the whole answer. A judge who sentences a serial killer to a prison sentence violates the interest of another human. So does a police officer who arrests somebody who is trying to shoot children in playground. But these acts aren't immoral. On the contrary, they're very moral.

In killing Smith, I'm undermining the integrity of social institutions on which the happiness and welfare of people in general depend. That's why what I'm doing is immoral. In killing the deranged gunman, the police officer is protecting those very institutions. That's why what he's doing is moral.

4.2 Why flourishing depends instrumentally and also constitutively on social institutions

Consider the aspirations that people in fact have. They want to be caretakers, heads of state, physicians, nurses, psychotherapists, engineers, historians, novelists, composers, artists, art-critics, literary critics, military officers, police-officers, lawyers, judges, and writers of philosophical treatises. All of these pursuits would be impossible were it not for social infrastructure. One reason for this is that nobody would live long enough to write any philosophical treatises or, for that matter, to educate a person for a long enough period to give that person the ability to write such treatises.

Another, deeper reason is that none of those aspirations would mean anything in a condition of lawlessness. Wanting to be a judge or a civil

engineer in a state of anarchy would be like wanting to be a father but to have no children. The existence of social structures is constitutive of personal identity. The aspirations that we identify with and that give our lives meaning are always defined in terms of social institutions. This is true even of the Browns of this world—the utterly selfish novelists (and philosophers, mathematicians, engineers, etc.). Brown's intention in writing novels isn't to do good; it isn't to consolidate existing social institutions. (He may even hope that his novels will stir up chaos and undermine all that is good in the social order.) But a world in which relations between people were categorically uncooperative would be a world in which novel writing would be meaningless. The value systems operative in such a world wouldn't validate novel writing. In such a world, writing novels would be like designing tea-cozies in the middle of a battle. Designing tea-cozies *does* have value—but not in that context. Novel-writing *does* have value—but not in the context of a war of all against all.

It might be said that, in such a context, novel writing *would* have value, but that, for obvious reasons, people would fail to see it—they'd be too absorbed in the business of surviving. While I grant that this position has merit, I would like to register a reservation that I have about it. People who write novels are, in so doing, reaching out to others. People do not, ultimately, write novels only for their own amusement. Of course, people sometimes write books that they know no one besides them will ever see. I myself have written two or three such books. But I wrote them because I knew that I needed to write them in order to develop the skills needed to write books that others might find worthwhile to read. The novelists that I know or have known of have all written a few early works that they never expected to publish. But they wrote them with the intention of writing other works that would be read by others.

It may be that Beethoven was a very self-involved person. But in composing music, he wasn't just talking to himself. He was addressing an audience. He may have thought that his contemporaries weren't the right audience—that it would take a few years for humanity to catch up to his music. But he was writing it in anticipation of some future audience that would understand it.

4.3 Narcissism and pseudo-immorality

There is no shortage of people who focus on their work and not only marginalize others but positively devalue them—of narcissists who see others as mere nuisances that have to be dealt with if they are to come up with their inventions or proofs, if they are to do their writing, their experiments, their art. But the efforts of such people are in fact addressed to an audience. It may be a future audience. It may even be an imaginary audience. But that imaginary audience will embody an outlook that is based on the outlook of actual flesh and blood people.

Not all dialogues are of a discursive kind. There are musical dialogues, architectural dialogues, technological dialogues, etc. When people meaningfully respond to other people's efforts, they are, in so doing, dialoguing with one another. Gianni Versace wasn't just designing clothes. His designs were meaningful responses to existing clothes styles. It's not as though his designs came out of the blue. Even if they were boldly innovative, they were responses to existing style. In fact, what constitutes bold innovation is entirely relative to what is currently out there. Beethoven's music wouldn't be boldly innovative today. It was boldly innovative because of what came before it and because of the evaluations of existing works implicit in it.

The same is true of novelists, inventors, economists—of anybody who is trying to "move the ball forward." There is a narrow sense in which such people may be deeply misanthropic. But their efforts are obviously connected—not just causally, but normatively, contensively—to those of others.

In a non-social world, those efforts themselves would not only be unrecognized but devoid of meaning. Of course, had Frege lived in such a world and, by some miracle, had arrived at and written the insights for which, in actuality, he is known, his *statements* would have been meaningful.[383] But his *saying* them would not have had any meaning. Given only that *what* one is saying is meaningful, it doesn't follow that one's *saying* it is meaningful. Suppose that you and I are emergency room doctors and we're talking about what we must do to heal patient X. I blurt out "I love peanut butter and chocolate ice cream." What I'm saying is meaningful—I'm reporting a truth about my psychology. But, relative to the circumstances, the act of my saying is meaningless; it has no place there; it's just a random

spark, a deracinated oddity. The same would be true of Frege's work in a non-social world. *What* he was saying would be meaningful; but his saying it would have no place there.

This shows that the dependence of human flourishing—on the morality of aspiration—on the morality of duty is not *merely* causal in nature. Of course, the former is indeed causally dependent on the latter. Flourishing would be a practical impossibility in a world where one's every moment was dedicated to saving one's skin. But there is also a non-causal, logical relationship between the two. The very idea of flourishing in such

a world is an incoherent one, at least to the considerable extent that flourishing involves participating in dialogues (of some kind or other) with others. And it's very hard to find a meaningful aspiration that doesn't involve this.

4.4 Why one's own welfare doesn't depend on one's complying with the morality of duty

But it doesn't follow that one must *oneself* comply with the morality of duty to have a flourishing life. Brown doesn't do so, but he has a flourishing life. It's necessary that it *be* complied with. If they weren't, Brown's literary endeavors would be drained of at least some of their meaning, for the reasons just discussed. But Brown doesn't necessarily have to be the one doing the complying. In writing brilliant novels in a world of otherwise decent people, a sociopath is doing something meaningful; he isn't in the same position that Einstein would have been in if he had been transported to the year 30,000 B.C.

Compliance with the morality of duty isn't sufficient for flourishing, and it's sometimes sufficient for *not* flourishing—for extirpating oneself. This was the case with Smith. He should have written novels, and left the charity work to many people who *want* to do it, instead of, like Smith, doing it because they feel they *have* to.

Thus, while there is a deep connection between morality and flourishing and, therefore, self-interest, that connection isn't as straightforward as some philosophers, and some non-philosophers, have alleged. And the *kind* of morality involved in flourishing—viz. the morality of aspiration—isn't the

kind that people usually think of when they hear the word "morality" (or some other comparable word). What people think of is the morality of *duty.* But the dependence of flourishing on the morality of duty is subordinate to its dependence on the morality of aspiration.

5.0 Moral integrity a requirement for psychological integrity

A person with no conscience rationalizes his failures, never taking responsibility for them and never, therefore, making the needed corrections to himself. For this reason, people of *no* conscience—so-called "psychopaths" or "sociopaths" (the terms are usually used interchangeably)—are dismal failures in life.

Before we continue, there are two things that must be made clear about the meaning of the word "psychopath." Psychopathy must be distinguished from psychosis. To be a psychopath is either to have no conscience or to have one that functions only intermittently and, on those occasions, only to a small degree. To be psychotic is to have a break with reality. Psychopaths aren't necessarily, or usually, psychotic or *vice versa.*

The other point is that being evil isn't the same thing as being a psychopath. A lot of deeply evil people have functioning consciences. Not all evil people are psychopaths (but most, if not all, psychopaths are evil). Evil people can, and usually do, have consciences.

There are two, very different reasons why this is so. And a discussion of these reasons will help make it clear why a certain morality (of duty) is inherent in one's succeeding in any non-trivial way and, therefore, in one's flourishing.

5.1 The first reason

People can be evil in some contexts but not others; and, in the contexts where they're evil, they're not always evil in *every* respect. Roy DeMeo was a Mafia Capo. He created and ran the largest car-stealing ring in New York City history. He also personally killed around 200 people. When he killed, it was usually for strictly business-related reasons; but sometimes it was purely for personal satisfaction. And he liked killing. Roy DeMeo was an evil

person. He made his living in a completely evil way.

But he was a very good father and husband. In his autobiography, his son Albert wrote that he "could not have asked for a better father." Albert never denies that his father, in his professional life, was evil. In fact, his entire autobiography is about how Roy was evil. More specifically, it's about how Roy DeMeo, while being such an evil person in one context, could be such a good person in another.

5.2 The second reason

This brings us to the second reason why evil people can have consciences. What Roy DeMeo did for a living was wicked in the extreme. But even in his professional life, Roy DeMeo conducted himself with a *certain* integrity. He didn't break his word to those with whom he was professionally transacting; he repaid his debts, honored his contracts, etc. (This is acknowledged by fellow partners in crime and also by the law enforcement officials who worked on his cases.)

DeMeo's compliance with these norms was, of course, largely prudential. A mafia captain can't break his word to his Mafia cohorts and expect to live long. But I don't think that's all that was going on. I think that, like most professionals, Roy DeMeo took pride in his work, and he thus wanted to do it properly. Though his reasons for doing it properly were *almost* entirely prudential, I suspect that a sliver of professional integrity was involved.

People, even Mafiosi, take pride in their work. Their professional behavior, though *largely* attributable to self-interest, isn't usually *entirely* so. (And when it is, their job-performance is noticeably substandard.) It is in this point that a genuine connection between the two kinds of morality is to be found; and this point is best illustrated by discussing the distinction between psychopaths and evil non-psychopaths (hence the preceding clarificatory points).

5.3 Psychopathy vs. evil

Non-psychopaths can be utterly evil. But there is a very big *psychological* difference between a psychopath and the non-psychopathic evil person. The difference lies in the fact in that, unlike the psychopath, the evil

nonpsychopath takes himself to task. When the evil non-psychopath falls short of his own benchmarks, despicable though those may be, he castigates himself. The psychopath does not castigate himself. He does not admit having done wrong. Instead, he rationalizes.[384]

Thus, a *certain* morality is inherent in the successful carrying out of anything of any scale, whether good, evil, or neither. One often hits roadblocks. When that happens, one has a choice: (i) take responsibility, and make the needed adjustments to one's own self; (ii) blame circumstances and admit defeat. If one can't admit that one has fallen short, one won't ever make the needed changes to oneself, and won't succeed in one's endeavors, be they morally wholesome ones or profoundly wicked ones.

Psychopaths are excellent swindlers and con-artists. Sometimes they end up with a non-trivial amount of wealth. But their accomplishments are trivial. To be sure, they can do a lot of harm. Psychopaths often swindle people and they sometimes kill them. But they seldom truly "make it big," even in the narrow sense in which Mao or a Stalin makes it big.

To get as far as he did, Mao had to evaluate his own conduct in a reasonably objective way. He had to ask himself, "Given that what I want is such and such, am I doing what I have to do, or am I falling short? And if I am falling short, how must I change myself to get back on track?" Mao had to censure himself and henceforth act in a way that wouldn't warrant additional self-censure.

To be sure, Mao was not censuring himself for what others would regard as moral lapses. Mao was quite at ease doing things that would have been unthinkable to most others. But there's all the difference in the world between somebody who self-censures for non-standard reasons and somebody who doesn't self-censure at all. And Mao did self-censure. And in so doing, his attitude was:

> (MM[385]) There are standards. I can fall short of them, just like anyone else. And I have fallen short of them. Like anyone else, I must change myself to make things right. Other people can't get themselves off the hook by blaming circumstances or by blaming the standards they've failed to meet. Neither can I. I must blame myself. I can't take refuge in excuses or rationalizations. I must cop to my failing and

change myself so that I don't repeat it.

In general, a certain morality of duty is inherent in any case of self-censure. Thus, embodied in any case of self-censure is compliance with a *certain* morality of duty. So far as one is censuring oneself, one's attitude is: "I'm just another person; and, just like others, I have to be judged for my failures." This attitude *is* the essence of the morality of duty.

Thus, a compliance with a *certain* morality of duty is inherent in any case of flourishing. Having a conscience is about holding oneself to standards; it's about stepping back from oneself, judging oneself dispassionately and, depending on the verdict, castigating or rewarding oneself accordingly. Psychopaths don't hold themselves to objective standards. When they fail, they don't blame themselves; they blame circumstances. When evil, non-psychopaths fail, they admit it and make the needed changes to themselves. Psychopaths are *so* self-indulgent that they don't do what they have to do to succeed. Unlike evil non-psychopaths, they don't even hold themselves to the nefarious ideals embedded in their projects, that being why their successes are confined to small cons and other such trivialities. But the successes of psychopaths—as distinguished from profoundly evil, non-psychopaths (e.g., Stalin)—are, in the big scheme of things, very minor. For this reason, history's great villains tend *not* to be psychopaths. Psychopaths are *so* bereft of morality that they don't have the self-discipline to get to positions of real power.

5.4 An objection to this line of thought

"But there's a very obvious point that you're missing," it will be said. "Although evil non-psychopaths do hold themselves to standards, no morality (of duty) is embodied in their doing so. Their self-censure and subsequent self-reforms are morally hollow and are of a strictly prudential nature."

Evil non-psychopaths are, psychologically if not ethically, in a very different category from psychopaths. And apart from the fact that they're evil, evil non-psychopaths are psychologically like non-evil, nonpsychopaths. Among the criteria for psychopathy are: (1) superficial charm; (2) inability to have a meaningful relationship; (3) lack of realistic life goals; (4) parasitic lifestyle; (5) manipulativeness; (6) pronounced tendency to rationalize

failure; (7) deceitfulness; (8) poorly integrated and shallow sexuality; (9) lack of remorse, shame, and guilt; and (10) an inability to operate within institutional structures.[386]

There are profoundly evil people (e.g., Roy Demeo) who have deep connections with others, who are capable of operating within institutional frameworks (e.g., Molotov), who don't have parasitic lifestyles (e.g., Ted Kaczynski), and so on. Demeo, Molotov, and Kaczynski were not psychopaths. They were evil nonpsychopaths. They had values. But their values happened not to be good ones.

Evil non-psychopaths typically have some of the nine traits just listed, but always lack most of them. Psychopaths, on the other hand, always have most or all of them. And non-psychopaths, when they have these traits, have them only in certain contexts and, in those contexts, in certain respects. Psychopaths, on the other hand, have these traits *generally*.

Also, non-psychopaths don't have them for the same reasons as psychopaths. Like psychopaths, evil nonpsychopaths, are deceitful and manipulative, but they aren't *generally* that way. There are usually people to whom they're honest. By contrast, psychopaths cannot be honest with anybody, including themselves.

Further, for non-psychopaths, the deceitfulness and manipulativeness are not ends unto themselves. But for psychopaths, they *are* such ends. Of course, they con for practical reasons. But they love to do it: it's what they do —it's their craft, their calling. There are non-psychopaths (e.g., DeMeo) who seem to feel no remorse or guilt. What's going on is that they do feel it, but not in the right contexts. DeMeo thought nothing of killing somebody for some a minor slight, but felt deeply ashamed over missing his son's baseball game.

It's been established that no one with a conscience has enough of these traits to merit a diagnosis of psychopathy, and also that anybody who doesn't have enough such traits does have a conscience. Given that conscience, and nothing besides, underlies all (non-coerced) compliance with the morality of duty, this shows that the self-censures of a Mao or a Genghis Khan are rooted in the very relationship to value and normativity that inhibits non-psychopathic, non-evil people from acting in antagonistic ways towards

others.

In conclusion, the self-discipline of a Mao is, from a psychological standpoint, no less conscience-driven than your reluctance to steal so and so's money. Both are rooted in a wish to adhere to a code of behavior that is larger than oneself and to which each human is supposedly accountable. Thus, the self-recriminations of an evil non-psychopath have a basis in a wish to comply with *a* morality of duty, albeit one that others might (claim to) hold in contempt. And given that the ability to self-censure, and change oneself in light of one's self-censures, underlies all cases of non-trivial flourishing, in both the evil and the good alike, it follows that a wish to comply with *a* morality of duty underlies, and is essential to, any case of flourishing.

6.0 How differences in people's views concerning flourishing mask underlying moral agreement

Not all people have the same moral values; not all cultures have the same moral values. Given this, it might seem as though I'm imposing my own idiosyncratic conception of morality on the rest of humanity by saying that excellence is flourishing and, consequently, that one is acting rightly insofar as, and only insofar as, one is acting in a flourishing-conducive way. What about people who believe that to be good is to obey God's will? Or those who identify goodness with fidelity to one's country? Or those who identify it with adherence to the Marine code of conduct? Or those who have some other view? Am I not simply dismissing these various different conceptions of morality, all of them reasonable, in saying that moral goodness is (the promotion of) flourishing?

In identifying goodness with flourishing and the promotion thereof, I am not mistaking my personal idiosyncrasies for moral universals; nor am I saying anything inconsistent with anyone's moral practices or beliefs. The differences in moral outlook between people and cultures have to do with differences in their views as to what *constitutes* flourishing.

Nobody thinks that flourishing is bad or that its frustration is good, but people disagree as to what it is to flourish. Some people think it involves connectedness to the ways of one's ancestors and, thus, in fidelity to tradition; others that it consists in breadth and intensity of awareness; others that it consists in having harmonious relations with others; others that it

consists in submission to God's will; others that it consists in loyalty to one's comrades; and so on.

A story about five people—Smith, Jones, Lewis, Brown, Clancy, and Ramsey—will substantiate these points. The one thing that Smith truly values, in both himself and others, is a keen intellect. The one thing that Jones truly values is wealth. He holds people who don't have it in contempt, and his one goal in life is to make more and more of it. The one thing that Lewis truly values is power—the ability to dominate others, to subjugate them to one's will. The one thing that Ramsey truly values is loyalty to country and to national tradition. The one thing that Clancy truly values is doing God's will. Finally, the one thing that Brown truly values is kindness. Money, power, and intelligence are fine; but whatever value they have is derivative of their ability to create a world in which people treat one another with decency and respect.

Each of these people realizes that, if the world is to function, other people must have values other than the one she happens to cherish. So, for example, Smith realizes that a world in which nobody valued anything other than a keen intellect would not be a good one. People would be too self-involved—they wouldn't worry about their children or their professional obligations, only their own intelligence quotients. People would be forced by economic necessity to become plumbers, hospital orderlies, etc. But their hearts wouldn't be in their work, and their performance would often be shoddy. Smith knows all this and thus concedes that things besides keenness of intellect have *a* kind of value—but the value in question is of a strictly instrumental nature. People must have other character traits—for example, they must have at least some regard for others, they must not neglect their children, they must in general (though obviously not always) keep their word, etc. But Smith thinks that such traits, though necessary, aren't in and of themselves, meritorious—they are good in an instrumental, not an intrinsic, sense. But for the fact that they may conduce to the sharpening of one's intellect, they are worthless. Each of the others believes the same thing *mutatis mutandis*.

These people disagree about what constitutes flourishing, but they all believe that flourishing is the one thing of value. No two people disagree that flourishing is good. At most, they disagree about what flourishing involves.

Chapter 23

Kant's Ethics and His Attempt to Identify Morality with Self-Interest

Immanuel Kant was a great epistemologist and also a great ethicist. Unfortunately, his insights are often embedded in confusion, and it isn't always easy to separate what is of value in his work from what isn't. This is especially so where his views on ethics are concerned, as we're about to see.

Like the present author, Kant held that it *is* categorically in one's interest to act ethically. But Kant's conception of morality bears little resemblance to the conception just described; and, partly for this reason, Kant's attempt to show that self-interest and morality coalesce bears little resemblance to the argument just presented.

To understand Kant's theory as to why morality and self-interest coincide, we must know what Kant believes morality to be. Kant puts forth two theories of morality.[387] According to the first, an act is moral just in case, were everyone else to act similarly, that act would not be self-defeating. According to the second, an act is moral to the extent that it embodies a respect for people's ***autonomy.*** Autonomy is self-determination, and for a person to be autonomous is for him, not someone else, to decide how he'll live. The opposite of autonomy is "heteronomy." Somebody is heteronomous to the extent that others decide how he lives. Thus, according to Kant's second ethical theory, autonomy-conducive acts are good and heteronomy-conducive acts are bad.

Kant believes these two positions to be equivalent. He's wrong. One can respect other people's autonomy without acting in such a way that, were everyone else to act similarly, one's act would be self-defeating. Because it wrongly assumes otherwise, Kant's attempt to link morality with self-interest fails.

1.1 Kant's two categorical imperatives

A ***categorical imperative*** is a rule that one should follow *no matter what.*

("Imperative" means "command" or "rule," and "categorical" means "unconditional.") Kant asked:

(ACI[388]) Are there any categorical imperatives?

He said "yes." Before we can say whether he was right, we must point out that there are three distinct questions that ACI can reasonably be seen as asking:

(1) Are there any moral rules that one simply *should* follow?

(2) Are there any rules of any kind at all that one simply *should* follow?

(3) Are there any moral rules that hold without exception? In other words, *supposing* that one's intention is to be moral, are there any moral rules that one must follow under any circumstances?

The answer to (1) seems to be "no." The statement:

(a) "one should refrain from committing immoral acts"

is ambiguous between:

(b) "one should act morally *if* one's objective is to act morally,"

and

(c) “one should act morally, even if it it’s in one’s interest *not* to do so.”

(a) is expresses a hypothetical imperative. (It’s also utterly vacant.) And (b) is clearly false.

In general, what one should do depends on what one’s intentions are. Therefore (1) is false. Similar reasoning shows that (2) is false.

Is (3) correct? Yes. One has a categorical moral obligation, I should think, to refrain from obliterating all life. And, of course, one has a categorical moral obligation to act morally.

But are there any categorical moral imperatives that, unlike the two just mentioned, *aren’t* utterly trivial? Kant said “yes.” And he identified what he believed to be two non-trivial categorical imperatives:

(CI_1[389]) An act is moral if it wouldn’t be self-defeating were everyone else to act similarly

and

(CI_2) An act is wrong if it’s a case of treating a person as an object, not a subject, and an act is right if it’s a case of treating a person as a subject not an object.

CI_2 doesn’t need much of an explanation. It’s obviously wrong to undermine another person’s autonomy. And, although it may well be right to undermine one person’s autonomy to preserve the autonomy of a million others, that doesn’t really redound to the discredit of CI_2.

The idea behind CI_2 is that, given any bad act x, if everybody did x, nobody could successfully pull x off. It be pointless to lie in a world where everybody lied, since nobody would believe anybody. It would be pointless to try to steal in a world where everyone was a thief, since everyone would go out of their way to make their possessions utterly theft-proof.

Kant thinks that CI_1 and CI_2 are equivalent. An act is one of using somebody else, Kant believes, if it would be self-defeating in a world where

everybody acted that way, and an act is one of *not* using somebody else if it wouldn't be self-defeating in such a world.

One can see why one would think this. In a world where everyone was constantly trying to con (or otherwise use) everyone, everyone would be on their guard, making it impossible, so one might infer, for anyone to con (or otherwise use) anyone.

But CI_1 and CI_2 are not equivalent. First of all, people *could* use people in a world where everybody was a user. In a world of con-men, the smart con-men could con the not so smart con-men.

In addition to showing that CI_1 and CI_2 are not equivalent, this shows that CI_1 is wrong. Though obviously morally bad, acts of con-artistry wouldn't necessarily be self-defeating in a world of con-artists. By the same token, there are morally good forms of conduct that, if engaged in universally, couldn't be successfully carried out. Physicians heal people; healing people is good. But nobody could successfully practice medicine in a world of physicians. Who would grow crops, build roads, enforce the law, and so on?

1.2 The fallacy embodied in Kant's search for a categorical imperative

A story will expose a confusion that vitiates Kant's ethical system. You are friends with Smith. Smith has money. You don't.

Your children need medication. If they don't get it right away, they'll die. Smith is a genuinely well meaning person who doesn't want your children to die. But (for some reason or other) it's impossible for him to lend you the money. The only way for you to get it is to steal it from him. There truly are no other alter

natives. You have exactly two choices: (1) Steal from Smith. (2) Let your children die. You choose to steal from Smith. The result: your children live, and Smith is bankrupt. You obviously did the right thing. Lost money can be regained. Lost life cannot. You don't regret making this choice. But you do regret *having* to make it. Smith is a good guy, and you really wish you hadn't had to steal from him.

Does this scenario show that there are exceptions to the principles that stealing is wrong? No.[390] Stealing from Smith is the lesser of two evils. It's obvious that, under the circumstances, the *right* thing to do was to steal the money. But in some cases the "right" thing to do is merely the *least wrong* thing to do; and this is one of those cases.

Thus, the principle that stealing is wrong *does* hold categorically. But this doesn't mean that you should never steal. What it means is that, although you have an obligation—a *categorical* obligation—not to steal, you have an even *stronger* obligation not to let your children die. A moral principle outweighed is not a moral principle extinguished. There are many categorical moral obligations. Lying, cheating, stealing, manipulating, wounding—these are all bad things. They are things that are categorically bad. But oftentimes these bad things are the lesser of two evils. Torture is wrong. But it's more wrong to allow a nuclear device to be detonated in the middle of a city. The circumstances being what they are, torturing one person may be far less wrong than the alternative (namely, letting many innocents die).

Suppose that, by virtue of the fact that you *had* to steal from Smith to save your children, you ceased to have *any* obligation not to steal from Smith. In that case, any remorse you felt as a result of stealing from him would have no objective basis. It would be the emotional counterpart of a hallucination. But, under the circumstances, it would be monstrous of you to feel *no* remorse at stealing from Smith. Of course, your choosing *not* to steal from him would have been far more monstrous, that being why, under the circumstances, the right thing to do was to steal. That said, it's appropriate that you feel *some* remorse at having depleted Smith's lifesavings. The reason for this is that, even though your obligation to save your children was far stronger than your obligation not to steal from Smith, the latter obligation did not cease to exist. It was merely outweighed. This suggests that the

obligation not to steal *does* hold categorically. In contexts where it appears not to hold, it *does* hold, but is outweighed by some other, stronger obligation. Presumably, the obligation not to kill innocent people is one that holds categorically, even though there are circumstances (e.g., ones where, in order to save a hundred innocent people, one must kill two innocent people) where it can be overridden.

Bearing these points in mind, let's reconsider Kant's two categorical imperatives:

(CI_1) Never commit an act that would be self-defeating if performed generally.

(CI_2) Always treat people as subjects, never as objects; in other words, respect other people's autonomy, without fail.

CI_1 entails that one should never lie—even if one must do so to save a hundred people.

CI_2 entails that one must never manipulate anyone—even if one must do so to prevent a nuclear bomb from going off in a crowded city.

The fact that Kant's ethical system has these consequences definitively shows that it is false. The nearest thing to Kant's ethical system that isn't obviously false is this:

(CI_2*) Although it's categorically wrong to treat people as objects, there are circumstances where one's obligation not to treat one person as an object is outweighed by some other obligation (e.g., by one's obligation not to allow ten million people to die).

1.3 Another Kantian blunder: his failure to distinguish the good from the praiseworthy

Kant held that nothing other than a "good will" has any moral value. By a "good will" he meant an intention to follow "the moral law." By "the moral law" Kant meant CI_2.

Here is Kant's reasoning:

> (KR) An intention to do good cannot possibly be bad. Of course, an intention to do good can have bad *effects.* Sincerely intending to save the children, you feed them X, which is lethal, genuinely believing that it's Y, which is nutritious. But *in and of itself* your intention to do good is utterly beyond reproach.
>
> But anything *other* than an intention to do good can be bad. A cool-headed, intelligent villain is that much more diabolical than a hysterical, unintelligent one. A wealthy, healthy villain is that much worse than a poor, sickly one. It is only *if* good people have them that intelligence, self-control, wealth, and health are good things. This means that those things are not *inherently* good.
>
> There is ultimately but one moral law, namely: respect other people's autonomy. It follows that anything other than an intention to treat people as subjects, as opposed to objects, lacks any inherent moral worth.

It might be that, unless lobotomized, a certain villain will use his formidable cognitive powers to do harm. But surely lobotomizing someone is, at best, the lesser of two evils. This shows that, although a villain's intelligence *is* inherently good, the existence of that inherent good may be incompatible with the some other, far greater inherent good. If, supposing that he isn't lobotomized, that criminal will kill millions of people, many of them intelligent, than one inherent good—saving the lives, and *a fortiori* the intellects, of millions of people—outweighs some other inherent good—preserving the intellect of some one person.

Imagine the following. Brown is a happy two year old infant. Brown didn't have to work for his happiness. He just happens to be happy. Question: Does Brown's happiness have moral value? Yes. Obviously. That's why it would be wrong to make Brown unhappy. Second question: Does Brown deserve *praise* for being happy? No. His happiness, though morally good, isn't *commendable.* Many things that are morally good are not commendable. This point is clarified and defended in Chapter 19, Sections 3.0–4.0.

In addition to being false, Kant's position is self-undermining. It would

make no sense at all to say (a) that autonomy isn't inherently good, while also saying (b) that an intention to respect autonomy *is* inherently good. So if Kant is right to think it inherently good to intend to respect autonomy, then autonomy must itself be inherently good. But in that case, Kant is *ipso facto* wrong to think that there is nothing inherently good other than an intention to respect autonomy. Which is not to mention that the intelligence of a villain is indistinguishable from his autonomy: take his intellect away, and you take away a good part of his autonomy, if not the entirety of it.

1.3.1 Another Kantian blunder: his failure to distinguish the good from the praiseworthy (continued)

A brief detour through the philosophy of law will help us identify the blunders in Kant's reasoning that underlie his ghastly view that nothing other than a good will has any inherent worth.

In his (1977) article "Is law a system of rules?" Ronald Dworkin (1931–) made the profound point that moral principles[1] have a "dimension of weight," whereas legal statutes do not. There aren't degrees of legality. An act is either legal or it isn't. Laws either hold or they don't. Of course, some laws are hedged with exception clauses. ("Killing is illegal—unless done in self-defense.") But given a duly hedged law, one is either in violation of it or one is not in violation of it. Legality is binary.

By contrast, morality is *not* binary. I have a professional and ethical responsibility to teach class on Tuesday. But I also promised a friend that I'd meet him for lunch. (When making that promise, I forgot that I had to teach.) I can't cancel the lunch-date. (I have no way of reaching my friend.) Presumably, my obligation to teach *outweighs* my obligation to keep my lunch-date with my friend. Each obligation has some weight; but one of them has more weight than the other. I still have *an* obligation to meet my friend. But I can't fulfill that obligation without shirking some greater obligation. But I still *have* that obligation. That's why, in standing my friend up, I'm behaving in a decidedly substandard way.

Kant evidently thought that, just like legal statutes, moral laws are binary:

a given law either holds or it doesn't. Kant was wrong. Moral laws, unlike legal statutes, can have varying degrees of relevance.

1.4 Kant's argument as to why morality and self-interest coalesce (KA)[391]

> (KA)We have value to the extent that we are rational. We're better than chipmunks and jellyfish to the extent that we are rational and they are not. (If chipmunks were to become rational and we were to become irrational, they'd have more value than we would.) Thus, so far as we subvert our own rationality, we undermine ourselves. Given an immoral act, the very idea of a world in which everyone committed that act is an incoherent one. It follows that immoral acts themselves are incoherent and, therefore, irrational. Since irrationality is self-undermining, and so is immorality; and it's thus inherently in one's interest to be moral.

This argument involves a massive non-sequitur. Kant says that the very idea of a world in which everybody lies and in which people still benefit from lying is an incoherent one. He concludes from this that in *our* world one is acting incoherently in lying. But this doesn't follow. Whether an act is rational in a given situation depends on how that situation is. Other situations are irrelevant. Whether an act is rational in our world depends on how our world is. In our world, a person's lying doesn't entail that everyone else lies; nor is a person's lying necessarily self-defeating for some other reason.

Kant was an exceptionally brilliant man. So why did he commit this ghastly logical faux pas? A story will help us answer this question. Smith is an auto mechanic who, being self-taught, relies entirely on rules of thumb. He doesn't know general principles of engineering and must deal with each car on a case by case basis. Given a car problem, Smith can't bring general principles to bear on it, and it is only by becoming intimately familiar with the idiosyncrasies of the car in question that he can figure out what is wrong with it.

Let's contrast Smith with that paragon of rationality, Isaac Newton. Newton could understand a certain phenomenon (e.g., a certain planet's orbiting in a certain manner around a certain star) in terms of general

principles. This was why Newton, unlike his predecessors, was able to see that the tides were to be explained in terms of the same physical principles as the trajectory of a cannonball and that both were to be explained in terms of the same principles as the orbiting of the Earth around the Sun. So far as he was able to see such inner similarities between apparently different situations, Newton was able to discern the real structure of the world and was, to that extent, genuinely rational. By the same token, in failing to understand particulars in terms of universals, Smith is failing to see the real structure of the world and is thus failing to be rational. This suggests that thought is rational to the extent that it is driven by a knowledge of principles that can be universalized—that is, that hold generally and aren't context-specific.

Given these very reasonable points, all of which the present author accepts, Kant infers that an act is rational to the extent that it is driven by principles that can be universalized. The principle that underlies a lie—viz. *I can tell a lie and benefit from it*—isn't one that can be universalized and for that reason, Kant believes, isn't rational. So by lying, thinks Kant, one is making oneself less like Newton and more like Smith, and one is therefore diminishing oneself.

So far as this argument appears probative, it is because the term "universalizable" is being used in an equivocal manner. Let us suppose that, from every standpoint other than a purely ethical one, my lying to Larry is the right move. I've figured out the relative weights of the various pros and cons, and factored them into an accurate and honest appraisal of my own objectives. Let's go even further and suppose that the reasoning that goes into my decision to tell this lie is driven by a thorough understanding of general principles. In this context, we are in effect supposing, I am a Newton, not a Smith.

Under these circumstances, the principle governing my act is universalizable in one sense and nonuniversalizable in another. It's non-universalizable in the sense that it *shouldn't* be universalized. But it *is* universalizable in the sense that, like Newton's reasoning and unlike Smith's, it's guided by an understanding of principles that hold generally. Given only what we've said, we have no choice but to see Kant as having utterly failed in his attempt to link morality and self-interest.

1.5 Why Kant's attempt to connect morality and self-interest is not a complete failure

But that attempt was by no means a failure. The principles driving that attempt are accurate and profound ones. He just wasn't putting those worthy principles to good use in this particular context. But Kant has real insight into the nature of rationality and autonomy and into their relationship to morality; and so far as there is a connection between rationality and autonomy, on the one hand, and morality, on the other, it is, I believe, to be understood in Kantian terms.

The essence of Kant's ethics is generally seen as lying in his contention that:

(CI_1[392]) An act is moral if it wouldn't be self-defeating were everyone else to act similarly.

CI_1 is false, and Kant's ethics is a failure if CI_1 is the core of it. But the real essence of Kant's ethics lies, not in CI_1, but in his contention that:

(CI_2*[393]) An act is good to the extent that it's autonomy-conducive and bad to the extent that it's heteronomy-conducive.

If it isn't outright correct, CI_2 is a very good approximation to the truth. It is morally acceptable to undermine a given person's autonomy *only* to the extent that doing so is necessary to prevent some greater violation of autonomy. Thus, we consider it morally appropriate to attack or kill somebody *only* to the extent that, by doing so, we're protecting somebody else's autonomy—for example, our own or that of our loved ones. And although we consider it morally acceptable to micromanage the lives of children and thus to encroach on their autonomy, we hold this *only* because we feel that, were we not to do this, our children would fail to develop their autonomy. They'd become undisciplined slobs who, having no self-control, lacked autonomy.

By the same token, protecting a given person's autonomy is wrong *only* to the extent that doing so leads to some greater violation of autonomy. We put people in jail and thus violate their autonomy. But so far as this is

justified, it's because, were we not to do so, those criminals, and those who followed in their footsteps, would be responsible for even greater violations of autonomy.

Thus, just as CI_2* says, an act is moral to the extent that it is autonomy-conductive and immoral to the extent that it is heteronomy-conducive. Given how intuitive CI_2* is, it's easy to see how an acceptance of it could be an ethicist's starting point; and given how well it fits the data, it's easy to see how it could be an ethicist's finishing point. But given how contrived CI_1 is, there's no real chance that it could be an ethicist's starting point; and given how wrong it is, it's unlikely to be an ethicist's finishing point.

Taken collectively, these points suggests that CI_2* *was* Kant's starting point and that, so far as he accepted CI_1, it wasn't really because he believed it, but because, for some reason internal to the mechanics of his theory, he felt he had to accept it. This analysis is substantiated by the fact that, despite his gargantuan intellect, saw CI_1 and CI_2* as being equivalent, even though they very clearly are not.

All of these facts fall into place when we remember that, in Kant's view, what gives us value is our rationality. Our rationality is what makes us better than animals. It is what makes us human. It is what makes us what we are. This suggests that, in Kant's view, rationality is the greatest good. How is this to be reconciled with Kant's belief that goodness is identical with autonomy-conduciveness and badness is identical with heteronomy-conduciveness?

1.6 Rationality vs. autonomy, practical vs. theoretical reason

The answer lies in the fact that Kant believes that rationality *is* autonomy. There is much to be said for this position, as we'll now see.

First of all, what do the terms "rationality" and "autonomy" mean? To be autonomous is to decide for oneself what to do and how to live. A creature that can't make *decisions* can't be autonomous. To be rational

is to regulate one's own thought processes in accordance with logical norms.

Thus, to be autonomous is to decide for oneself how to live, and to be rational is to decide for oneself how to think.

"But couldn't one be *irrational,*" one might ask, "and still decide for oneself how to think? Couldn't one learn all the relevant logical norms, and make a decision *not* to comply with them in one's thinking? To be sure, it isn't likely that one would do this. But isn't it at least a theoretical possibility? And given that it is, doesn't it follow that your definition of "rationality" is wrong?"

In order to carry out such a program, one's thinking would have to be rational. In any given situation, one would, by the objector's own terms, (a) have to figure out what the rational thing to think was, and (b) decide *not* to think it. And the first part, at least, would involve being rational. The second part, admittedly, wouldn't be rational—if one knows that the right conclusion to draw is Q, but one infers not-Q instead, one isn't being rational. But in the case of the person just described, that irrationality would be parasitic on an existing rationality.

Also, if somebody *knowingly* draws the wrong inference, one is "irrational" in a different way from somebody who fails to see what the right inference is to begin with. The second sort of irrationality is purely cognitive. The first has to do with the relationship between cognition and action. The one has to do with what philosophers sometimes refer to as theoretical reason; the other has to do with what philosophers call *practical reason.* Theoretical reason has to do only with the drawing of inferences; it has nothing to do with the choices that one makes on the basis of one's inferences. The degree to which one is "theoretically reasonable" is a function only of the degree to which one draws the right inferences, it being irrelevant what one decides to do on the basis of those inferences. Practical reason has to do with the relationship between the inferences one draws, on the one hand, and the intentions in which they eventuate, on the other.[394]

Given this, suppose that Smith decides to live the sort of life described by the objector. In other words, Smith has resolved that he'll always conclude the opposite of what the norms of good reasoning require him to conclude. This means that, in any given context, he'll (a) figure out what, given the relevant data, the right conclusion to draw is, and (b) having figured that out,

he'll take it upon himself to draw the opposite inference.

Smith is a paragon of a theoretically reasonable person: by hypothesis, Smith's reasoning is impeccable, at least to the extent that he's successful in his efforts to carry out his agenda. So far as Smith is *irrational,* it's because of what he decides to *do* with the correct inferences that he's irrational. To be sure, what he decides to do with those inferences is have a belief—it isn't to do anything overt, like go for a bike ride or have a beer. But it's still, in the relevant sense, an action. So Smith *is* rational, at least insofar as rationality has to do with reasoning. So far as he's irrational, it is in the practical, not the theoretical, sense. Thus, we don't have in Smith a counterexample to the principle that to be rational is to decide for oneself how to think.

"Very well," one might retort. "Contrast Smith's case with Brown's. Brown skips the first part of Smith's two-step approach. Brown resolves one day that, in any given context, he *won't* figure out what the right inference to draw is. Smith *first figures out* what he, given the available data along with the relevant logical norms, ought to believe. Having done that, he resolves to believe the opposite. But Brown skips the first part. He has resolved just to let his beliefs arise spontaneously and, therefore, not to be responsive, even in a negative way, to the demands of canons of logic. It's a datum that, so far as he carries out this plan, Brown's thoughts aren't rational. (They may be *correct.* But, being entirely reflexive, they aren't *rational.*) But it's also a datum that Brown has decided *for himself* how to think. This straightforwardly shows that you are wrong to say that to be rational is to decide for oneself how to think.

But Brown has decided for himself what to think in the sense that he's decided not to decide for himself what to think. He has decided that, henceforth, he won't decide for himself what to think. Once he gives himself over to that life-course, he is *ex hypothesi* no longer rational. Therefore, we don't have in Brown a real counterexample to our analysis.

Basically, deciding for oneself how to think *always* involves regulating one's thoughts in accordance with norms. Even if one picks the wrong norms —even if one decides not to go with ***modus ponens, modus tollens,*** etc. and to go with other, spurious forms of inference—one's correctly *applying* these spurious forms of inference *itself* involves one's having non-spurious beliefs

as to how they apply to the data at one's disposal. So given somebody who decides to think in accordance with the wrong principles, he can carry out his agenda *only* to the extent that he correctly sees how those principles bear on the data and, therefore, only to the extent that he is rational.

No matter how you slice it, rationality is *intellectual* autonomy. In other words, to be rational is to decide for oneself how to think.

Let's compare rationality with autonomy. To be autonomous, I have said, is to live as one chooses. One's being autonomous, I propose, is the same thing as one's deciding what to do on the basis of norms that one has adopted. A good way to see why this is (I believe) the right approach is to consider an obvious objection to it:

(NSG[395]) Your analysis is definitely wrong. Consider the case of Williams. Williams has decided that he'll do whatever he wants to do. He won't examine his actions, or the intentions behind them, in light of norms. He'll just act. But he's autonomous. That's a datum. After all, he's deciding for himself how to live.

Here's why this is off the mark. To the extent that Smith carries out his agenda, he isn't *deciding* anything at all. True—it is indeed a decision on his part that initiates his entry into a life of the sort described (viz. one in which he just does what he wants, never bothering to assess the consistency of his desires with norms of any sort). But once he's in that life, he isn't *choosing* anything at all. A creature that doesn't deliberate doesn't *decide* anything. It just acts. And a creature that deliberates is *ipso facto* one that is trying to act on intentions that are in compliance with some set of norms.

Do dogs make *choices?*[396] Obviously dogs have intentions. (The dog intends to get to the ball before the other dog. That's why it's running so frantically.) But it seems wrong to say that dogs "choose" (or "decide"). Dogs don't deliberate. They just react. The dog sees red meat: he pounces on it. He sees his brutal master: he cringes. The dog's life is a series of programmed and therefore compulsive reactions to immediate stimuli. The dog doesn't call the shots; these hardwired reaction protocols call the shots for him, and it therefore doesn't make choices. Only creatures that deliberate

make choices. Sub-rational[397] creatures don't make choices. Non-rational creatures don't act; they *re*act.

Rational creatures don't just react. Their responses to situations are considered. That doesn't mean that they *deliberate* before every act. (Nor should they, since a life spent deliberating would be one of "analysis paralysis"—of over-thinking what one should do and therefore never doing anything. It's therefore irrational to deliberate too much. It's irrational to be too rational![398]) But their responses to situations are guided by *choices* that they've made, and *those* choices result from prior deliberations on their part. I see an ice-cold beer in the fridge. Were I to react instinctively, I'd drink it. But long ago I made a choice to live a certain way (to be an author and, therefore, to avoid things, such as getting drunk in the morning, that would make it hard for me to ply my trade), and my instantaneous decision not to drink it reflects that choice. So even though I don't deliberate, my act (of refraining from drinking the beer) is the embodiment of prior deliberation on its part. The dog's reactions, by contrast, do not reflect previously made life choices.

One last point before we close the argument: given a set of norms that one has decided to comply with, it's not always obvious how to comply with them. In such cases, thought is needed to know how it is that, given those norms, one ought to act in a given situation if one is to comply with those norms. And even in cases where it *is* obvious how to comply with those norms, it's obviously only to the extent that the right *judgments* come naturally to one. Thus, a prerequisite for practical rationality—that is, for autonomy—is theoretical rationality.

We can now say how rationality and autonomy are related. To be rational is to regulate one's thinking in accordance with norms. To be autonomous is to regulate one's actions in accordance with norms. The latter is a special case of the former, since, as we just saw, one's thinking must be appropriately norm-driven if one is to know how which actions on one's part will be incompliance with the norms of conduct one has adopted. Thus, rationality and autonomy *are the same thing*—or, rather, autonomy is a special case of rationality. Rationality is intellectual autonomy, and autonomy is practical rationality. So rationality *is* autonomy. A creature can

make decisions only to the extent that it is autonomous and, therefore, rational. (Non-rational/nonautonomous creatures have intentions. But they don't make *choices.* They don't deliberate. They just react.)

A creature that doesn't make decisions isn't an *agent,* at least not in the sense in which you and I are. *It* doesn't form its own intentions. It's intentions *result* from a convergence of factors (its innate and acquired psychological and physiological structure, its past experiences, the present context). But *it* doesn't form them. Its intentions are formed in an impersonal way. An autonomous being forms its own intentions. So an autonomous being is responsible for its intentions and, therefore, for the acts to which they lead. By contrast, a sub-rational, non-autonomous creature is responsible only for its actions. The intentions of a non-autonomous creatures are not the results of *agency.* Agency, for the non-autonomous, begins with intentions, but doesn't itself generate them. For the autonomous, on the other hand, agency is itself responsible for intentions. Words like "agent," "agency," etc. are ambiguous; and words like "subject" (as in, "he is a human subject") are correspondingly ambiguous. In one sense, dogs and infants are agents. In another, they are not. Which is why we don't hold 3-year olds criminally responsible for their acts. Two-year olds are agents in *a* sense of the word, but not in the same, robust sense in which you and I are agents. Words like "subject" (as in, "he is a human subject") are correspondingly ambiguous. For the reasons just given, agency in this more robust sense is identical with rationality and, therefore, with autonomy; and agency in this sense is identical with moral responsibility. Respect for a person's autonomy is therefore nothing other than respect for his agency and thus for his very existence, to the extent as he is, in the robust sense of the word, a subject.

Contrary to what Kant tries to show, none of this entails that one's own welfare necessarily involves one's respecting *other* people's autonomy (rationality/agency/subjecthood). But it shows that one's welfare depends on, if it isn't identical with, one's being moral to oneself—on one's not encroaching on one's own autonomy. It also shows that ethical responsibility and agency are coeval with rationality and autonomy. In any case, Kant's attempt to identify morality with self-interest, though not itself a success, is driven by a cogent analysis of the nature of agency and its connection to rationality and morality.

Chapter 24

Hedonism, Egoism, Utilitarianism, and Deontology[399]

1.0 Hedonism

There are two different doctrines that go by the name of "hedonism." One is an ethical doctrine about what we *should* do. The other is a psychological doctrine about what we *in fact* do.

According to ethical hedonism, people *have an obligation* to pursue their own pleasure and they have no other obligations.

According to *psychological* hedonism, human beings always seek pleasure. In other words, it is a law of psychology that, whenever a person acts, it is in order to experience pleasure.

Nobody denies that human beings can deliberately forego pleasure. But, according to the psychological hedonist, people deliberately forego pleasures only if they think it necessary to open up opportunities to experience pleasures greater than the ones they're foregoing.

A person can coherently accept ethical hedonism without accepting psychological hedonism. In fact, it would be incoherent to advocate ethical hedonism *unless* psychological hedonism were false. If psychological hedonism is correct, people will seek pleasure no matter what, and in that there is thus no point in *telling* people to pursue their own pleasure.[400]

1.1 Psychological and ethical hedonism evaluated

There is a difference between doing something with the *intention* of bringing about X, on the one hand, and doing something *knowing that X will come about as a result,* on the other. A physician has to perform a lifesaving emergency procedure, but there's no anesthetic. The physician *knows* that, by performing the procedure, he'll cause the patient to experience pain. But he doesn't do so *in order to* inflict pain. We know this, because if anesthetic were to become available, he'd use it.[401] Obviously the physician is in a

very different moral category from the physician who has access to anesthetic but, out of sadism, chooses not to use it.

So given only that Smith (who, let us suppose, is a latter day Mozart) *knows* that composing music brings him pleasure, it doesn't follow that he does so *in order* to bring himself pleasure. Nor is it the case that his primary reason for composing is to bring himself pleasure. He composes music because he sees that it has value; and *because* it has value, doing it brings him pleasure.

Unless Smith *valued* the act of composing, it *wouldn't* bring him pleasure. One can derive pleasure from composing only if one sees value in it. If one thought that composing was a waste of time, one wouldn't like it. (There may be people who, in the grips of a morality they don't really believe in, *tell* themselves that the pursuits they in fact value *don't* have value. But they're lying to themselves. We discussed this in the previous chapter.)

People don't play soccer (or do medical research or write treatises) *in order* to bring themselves joy. They do it because they value it, and because they value it, doing it brings them pleasure. The pleasure they experience in what they're doing is a reflection of their belief that what they're doing has value.

To be sure, there are cases where people act with the objective of experiencing pleasure. People drink alcohol and do drugs *in order to feel pleasure.*[402] People take vacations and watch TV *in order to experience pleasure.* I myself like to watch TV, but I don't do it because I think it has value. (It doesn't.[403]) I do it because I like to do it. But notice that it is precisely these acts that we tend to regard as base. Nobody thinks ill of Beethoven for spending so much time working on his music. Even though he enjoyed almost every minute of it, we don't see him as being in the same category as party boy Jones spends, who spends his nights doing cocaine and his days sleeping off the effects. We admire Beethoven (he was "dedicated to his craft"), whereas we revile Jones (he's "empty and dissolute").

Why do we have such different attitudes towards these two? It was in order to do something that he valued that Beethoven composed. It wasn't *in*

order to make himself feel good that Beethoven composed; and so far as composing brought him joy, it was because what he was doing had value. But it isn't to do something of value that Jones does cocaine. It is in order to have a good time that Jones does cocaine. So the joy that Beethoven felt was valuable, whereas the joy that Smith feels is not.[404]

It's obvious that pleasure has great value. But we've seen that there is *an* important kind of value that it lacks and ethical hedonism must be rejected for this reason. Ethical hedonism would say that Beethoven's moral obligation was to have a good time. Since it wasn't Beethoven's one intention in doing his important work to have a good time, and since his life obviously would have been eviscerated had he been forced to give it up, ethical hedonism requires the Beethovens of this world to *stop doing what they love.* But surely no viable ethical doctrine would require that.

Looked at superficially, ethical hedonism seems like a liberating, care-free doctrine. Looked at less superficially, it's a dehumanizing and cruel doctrine, as it requires titans like Beethoven to become munchkins like Jones.

1.2 A variation of the argument just given

Here's a brief recap of what we just said. Mother Teresa helped the poor because she saw value in it, not because she thought that she'd feel pleasure as a result of doing so. Because she saw value in feeding the poor, it gave her pleasure to do so. But that wasn't why she did it.

A story will make it clear why this claim is correct. X doesn't like people very much. To the extent that he has any feelings, whether positive or negative, about anyone, they're negative. He finds the suffering of others mildly amusing. But he doesn't much care whether people suffer or not. X undergoes cranial-surgery. (He has a tumor that has to be removed.) The operation goes well; X is healthy, physically and psychologically. X has the same attitudes he used to have—and also the same beliefs, memories, skills, etc.

But the operation has one curious side-effect. When X gives money to poor people (e.g., pan-handlers), he experiences exquisite pleasure as a result. The pleasure he experiences is *not* the pleasure that Mother Teresa experiences. It isn't his belief that he's making the world a better place that

leads to his feeling good. Giving money to poor people causes him to feel good in much the same way that shooting heroin causes somebody to feel good. The relationship is strictly causal. The difference between X and the heroin-user has to do only with the mechanisms involved. The one uses a syringe filled with heroin. The other uses bits of green paper.

So far as Mother Teresa derives pleasure from helping the poor, it's because she values it. She isn't doing it in order to feel good. He values it *only* to the extent that it brings him pleasure. That's why, if he thought he could experience comparable pleasure by hurting the poor, he would not hesitate to do so. But, if in a similar situation, Mother Teresa surely would hesitate to do so.

A similar story will help us appreciate the scope of this point. Y thinks that composing is a big waste of time and doesn't even like music. But he happens to have a knack for it. He doesn't like doing it. But *if* he does it for a little while, he experiences a huge surge of pleasure. Y's relationship to composing is like a heroin-user's relation to heroin. The only difference is that, whereas the one uses as syringe, the other uses a pen (or whatnot). Composing brought joy to Beethoven—but not for the reason that it brings joy to Y. It wasn't because of what it could do *for him* that Beethoven liked to compose. He valued it; and that's why composing made him feel good.

2.0 Egoism

Like the word "hedonism," the word "egoism" is ambiguous, and the two words are ambiguous in parallel ways. "Egoism" sometimes refers to a psychological doctrine, and it sometimes refers to an ethical doctrine.

According to ethical egoism, one ought never to act in a manner that isn't self-interested. According to psychological egoism, one can't act in a manner that isn't self-interested. It is one's sole intention in performing any given act to benefit from that act. The expected benefit may be simple pleasure or it may be an authentic enhancement of self.

The problems with psychological egoism are similar to the problems with psychological hedonism. It was in Beethoven's self-interest to compose music only because composing music *itself* has value. If Beethoven *didn't* recognize the value of composing music, he wouldn't have seen it as being in

his interest to do so.

Although we're obviously egocentric, our egocentrism isn't as morally empty as we're inclined to think. If you don't see value in being on the Supreme Court, being on it wouldn't mean anything to you, and you wouldn't see it as being in your interest to be on it. Thus, psychological egoism isn't a viable doctrine.

Also, there are well-known counterexamples to psychological egoism. There are people (e.g., Mother Teresa) who don't seem to be acting in their own interest, and *do* seem to be acting almost entirely in the interests of others. Advocates of psychological egoism deal with such cases by saying that Mother Teresa did these things because they made her feel good and that in doing them she was therefore acting in a purely self-interested fashion.

This position isn't tenable. Mother Teresa wouldn't have enjoyed doing these things unless she saw them as valuable. She did them because she saw that they were valuable, and because they were valuable, her doing them made her feel good. But that is very different from saying that she did them *in order* to feel good. (See the previous section for clarification of this point.)

What we said a moment ago concerning Mother Teresa shows that ethical egoism is false. She helped people *not* because it was in her interest, but because she saw value in it.

Of course, *given* that she structured her life around that value, it *came* to be in her interest to help people. And had Mother Teresa gone into investment-banking, it would have *become* in her interest to become a good investment-banker. But it obviously wasn't *ab initio* in her interest to go into investment-banking; nor was it *ab initio* in her interest to become a charity worker.

Are there any cases where it is in one's interest to do something that has no value *other* than the fact that it is in one's interest? By definition, any such cases wouldn't involve acts that involve the recognition that anything *other* than one's self-interest has any value, and such acts would therefore be ones that we are inclined to regard as enjoyable but meaningless (e.g., shooting heroin). But, although it's obviously in one's interest to have a meaningless good time now and then, it's hard to see how it could be in one's interest to do nothing other than that. It's very obvious, in fact, that one would be

diminishing oneself in an extreme way were one to do nothing but shoot heroin. Ethical egoism is thus an incoherent doctrine.

In the previous chapter, we argued that *a* kind of morality is embedded in any act that it is genuinely in one's interest to perform. We also argued that subject hood presupposes sensitivity to norms. What we just said about ethical egoism corroborates these points.

3.0 Utilitarianism

According to ***utilitarianism*****,** an act is good to the extent that it promotes human welfare and bad to the

extent that it undermines it.[405]

This seems like an utterly reasonable doctrine. But there are some well-known problems with utilitarianism, and many ethicists reject utilitarianism because of them. Some stories will help identify two of these problems.

Story #1: You are a physician. To save five unhealthy people who need organ transplants, you must either kill or vivisect four other perfectly healthy people and transplant their organs into the other five. (At least one of the unlucky four will donate more than one organ of his.)

Question: Should you kill the four to save the five? Presumably not. It seems wrong—a heinous violation of the autonomy of the four.

Story #2: A terrible murder has occurred. Somebody killed an innocent child in a grisly way. The public wants justice. The police happen to know that the real killer fled to a country from which he couldn't possibly be extradited without causing a war that would take many lives. Nobody else knows this. The police also know that if *somebody* isn't convicted of this crime and made to pay for it, the public will riot, the effects of which will be devastating. The police know that, were they to frame an innocent person, the public would be sated and these disasters would be headed off.

Question: Should the police frame an innocent person? Presumably not (for the same reason as last time). These two stories suggest that there are cases where, no matter how much good *comes* of some act, that act is still *just wrong.* If these two stories are representative, acts that fall into this category involve cases of autonomy violation. This, of course, substantiates Kant's

position that goodness = autonomy promotion and that badness = heteronomy promotion. Kant was a virulent opponent of utilitarianism.

There are other problems with utilitarianism. I'll focus on what I think are the most important ones. First, if everyone were always trying to maximize the welfare of others, everybody would be miserable, since no one would be doing what he wanted. Second, utilitarianism says that anything that isn't an act of trying to help everyone has no value and that an act has value only to the extent that it helps others. If this is right, then reading a book or listening to a good piano sonata has value *only* to the extent that it helps others. Third, utilitarianism can't accommodate the fact that personal loyalties have moral value. Given a choice between saving a stranger and saving your devoted spouse of 50 years, you should save the stranger if, by doing so, you stand to make people even infinitesimally better off than you do by saving your wife. But that seems positively monstrous. Fourth, if everybody were trying to make the world a better place with every single act, the world would quickly become a very bad place. For example, if judges broke the rules in order to serve the interests of justice, the legal system would collapse, and all the protections it provides would be withdrawn.

3.1 Rule utilitarianism

Helping people is good and hurting them is bad. This is the idea underlying utilitarianism, and it's obviously a good one. So many want to hold onto utilitarianism, despite the problems with it just discussed, and they typically deal with these problems by advocating a doctrine called "rule-utilitarianism" (RU).

According to rule-utilitarianism, one should identify a set of rules that, if generally complied with, lead to more happiness, and having identified those rules, one should comply with them. Physicians shouldn't vivisect people to save others. For, if they did so, everyone would walk in mortal fear of being vivisected, and life would become a horror. Physicians should identify a set of rules that, if generally complied with, lead to an optimal outcome. Among those rules, there is presumably one to the effect that you shouldn't vivisect or kill one person to save two.

Police officers shouldn't frame people every time doing so conduced to the public weal. For, if they did so, the framing of innocent people would be

routine, and the police would be our tormentors, not our protectors. They should identify and comply with welfare-optimizing rules. Among those rules, there is one to the effect that you shouldn't frame the innocent, even if doing so helps more people than it hurts.

People shouldn't be asked to sacrifice their personal ambitions or relationships for the sake of the greater happiness. For, if people did make such sacrifices, they would be unfulfilled—every would-be composer or novelist or mathematician would be stuck in some soup kitchen. And personal loyalties would be unsustainable, since they'd be so often violated, and life would be drained of all the good they bring. Thus, people should identify and comply with welfare-optimizing rules; and among such rules, there is presumably one to the effect that you shouldn't sacrifice your personal objectives or spurn your close friends.

3.1.1 Rule utilitarianism evaluated

RU is a reasonable doctrine, and I think it gives us insight into some otherwise opaque ethical issues. It's said, for example, that it's inherently wrong of nations to "violate the sovereignty" of other nations. But this doesn't seem true. First of all, sovereignty is a property, not of the people who live in a country, but of the government of that country. A government is legitimate only to the extent that it serves the interests of its constituents. If the government of some nation torments its own subjects, it would be wrong *not* to overthrow it, other things being equal. (The "other things being equal" hedge is important. Overthrowing a government tends to

do a lot of damage, since governments, even bad ones, keep order. Unless the damage that a government does exceeds the damage done by overthrowing it, it isn't to be overthrown. And for a government to reach that threshold, it must be exceedingly bad.)

But, so it could be argued, the world would plunge into never-ending war if countries were allowed to save other nations from themselves by invading them and destroying their governments.[406] So such vigilantism, even when just, must be prohibited. Points of this kind are easily abbreviated with the help of expressions like "a nation's inviolable right to self-govern."

Much talk of "civil liberties" is to be understood in rule-utilitarian terms.

Is it inherently wrong for a police officer to plant evidence on somebody who he knows to be a serial killer? Does that serial killer really have any of the "inviolable rights" listed in the Bill of Rights? No, and it wouldn't be a miscarriage of justice to bend the rules a bit to put a homicidal maniac behind bars. *But* were such rule-bending permitted, bona fide miscarriages of justice would ensue. (For example, the police would start framing innocent people they didn't like. And, in any case, the behavior of law-enforcement personnel would become unpredictable. And the fact that it was unpredictable would lead to regressive social changes of many kinds.) So when it is said that this or that individual has these or those "inviolable civil liberties," what is really being said doesn't concern that individual at all, and is instead to the effect that certain practices that are acceptable on this or that occasion would have bad consequences if they were to become the norm.

3.1.2 Rule utilitarianism evaluated (continued)

But despite its merits, RU isn't quite satisfactory. If RU is right, a person's right not to be vivisected is contingent on the fact that vivisecting him might lead to the vivisecting of many others and thus to rampant unhappiness. But surely vivisecting a person would be wrong *even if* it didn't lead to the vivisecting of others. Preventing Beethoven from composing music would be wrong even it didn't lead to other, comparable prohibitions. Preventing people from giving special attention to their friends and loved ones would be wrong even if doing so didn't have unwanted ripple effects.

Composing music is something that has *intrinsic* value. It's a good unto itself. But RU can't register this fact. To the extent that RU can recognize any value in Beethoven's work, it is of a purely instrumental kind. So far as that work has value, says RU, it lies in the fact that thwarting might set off a ripple effect which would lead to mass-unhappiness. SRU is systemically incapable of seeing the value of Beethoven's aspirations for what it is; it's systemically incapable of conceding any sort of intrinsic value to most of the things that we think to have such value (relationships, personal goals, and life itself). Thus, RU is false.

Composing music is something that has *intrinsic* value. It's a good unto itself. But RU is inconsistent with this fact. According to RU, Beethoven's musical activity has value *only* to the extent that preventing him from

engaging in it might set off a ripple effect that would lead to mass-unhappiness. RU is blind to what is *actually* of value in Beethoven's work. It's similarly blind to what is *actually* of value in being healthy (and, therefore, in not being vivisected) and in having meaningful relationships. Thus, RU is false.

4.0 The deontological approach

Utilitarianism is a form of ***consequentialism.*** This is the doctrine for an act to be good is for it to have certain effects. An act is right, says utilitarianism, if it maximizes human welfare.

According to ***deontological*** theories of ethics, for an act to be morally right is for it to be in compliance with the agent's duties and obligations to others. An example of such a theory is Kant's contention that:

(CI_2) An act is wrong if it's a case of treating a person as an object, not a subject, and an act is right if it's a case of treating a person as a subject not an object.[407]

If CI_2 is right, what makes an act right is not that it has certain effects, but that it embodies a certain respect for others.

According to the consequentialist, the ends justify the means. Anything that makes the world a better place is justified, says the consequentialist. If burning innocent children alive is what does the trick, then do it.

The deontologist rejects this. Those children have rights. The are ends unto themselves. They aren't to be used—even if the use to which they are put is the betterment of the human condition.

While reading what follows, it's important to bear in mind that utilitarianism isn't the only form of consequentialism. Some of the just-stated anti-utilitarian arguments are persuasive, and I am convinced by them. It's *inherently* wrong to prevent Beethoven from composing. The wrongness of it isn't contingent on some presumed ripple effect. But this shows only that utilitarianism is wrong, not that all forms of consequentialism are.[408]

4.1 Limitations of the deontological approach

In low-stakes situations, the deontological approach seems reasonable. You promised to meet an old friend for lunch. For whatever reason, you'd be doing the world slightly more good by breaking that date than by keeping it. But you should still keep it. A promise is a promise.

But when the stakes are high, consequentialism is clearly the right doctrine. If you have to break the lunch date to save a life, you should break it. If you have to kill one person to save a million, you should kill that person.

This suggests that, to the extent that it's distinct from consequentialism, the deontological approach is incoherent. By letting a million people perish to save one, you're violating *their* rights. The utilitarian approach doesn't lead to such gross rights violations and is thus, after a fashion, more deontological than deontology itself!

4.2 Rule utilitarianism revisited

Now that we've discussed the deontological approach, we can identify a problem with RU that we weren't in a position to discuss earlier. Depending on how it's interpreted, RU collapses into either act-utilitarianism or deontology. There will inevitably be cases where one should break rules that are otherwise worth obeying. There will be times when the evidence must be planted, when lies must be told, when rights must be violated, when confidences must be broken. Planting evidence to convict a serial killer isn't good—but it's not as bad as what the serial killer will do if he isn't convicted. What is the advocate of RU to say about such situations? If he says that you *mustn't* plant the evidence, no matter how much badness it will prevent, then he isn't really a utilitarian any more: he's saying that the rules must be obeyed for their own sake. He's saying that you must obey the rules *because it's the right thing to do.* But that isn't utilitarianism; it's deontology. If someone says that you *should* break the rules, then he isn't a *rule*-utilitarian anymore; he's an act-utilitarian.

5.0 A unified approach[409]

The deontologist doesn't want rights to be violated. The consequentialist wants to ensure a maximally good outcome. Really, then, they both want the same thing, since a maximally good outcome *is* one in which rights aren't

violated. Thus, they both believe it to be one's moral obligation to respect the rights of others. But they have different views as to how one should attempt to fulfill this obligation.

The consequentialist says:

> Don't just follow rules blindly. Do the right thing. *You* should make sure that people who need organ transplants get them. Don't be a faceless bureaucrat who hides behind the rules. Don't describe your cowardice and amorality as "respect" for people's rights. Don't pass the buck. Don't wait for "the other guy" to step up. Be a hero, not a zero.[410]

The deontologist says: Other things being equal, you'll do more good than bad by respecting people's rights and fulfilling your obligations to them, and you'll do more bad than good by violating people's rights and shirking your obligations to them. Stalin, Mao, and Pol Pot came into power because well-intentioned people decided that it was their lot to see to it that justice was done, the rules be damned.[411] The road to hell is paved with good intentions. So do your duty. Don't play God.

The differences in viewpoint between the consequentialist and the deontologist concern how to create a moral world, not what a moral world is; and they therefore don't disagree with each other as to the nature of morality.

The deontologist's position, as represented here, is similar to that of the rule-utilitarian, and the consequentialist's position is similar to that of the act-utilitarian. Given that utilitarianism is generally seen as a paradigm case of a consequentialist doctrine, it might seem that I must have misrepresented deontology. But the right conclusion to draw, I would submit, is that, just as I suggested, the deontologist and consequentialist don't disagree as much as they think they do.

Chapter 25

Religion

1.0 What does philosophy have to teach us about religion?

Nothing. The so-called philosophy of religion doesn't concern religion. It concerns puzzles that can be illustrated in connection with religious questions but are really of a strictly logical nature. Hise is one such puzzle:

> (OG[412]) Many hold that God is omnipotent—that, in other words, He can do anything. Let's suppose that this is true. In that case, God can create some task that He cannot perform. But if He can do that, He's not omnipotent.

OG doesn't have anything to do with God. It has to do with subtleties of logic that, although they can be illustrated in connection with statements concerning God, have nothing to do with His. We know this because OG perfectly parallels the following argument.

> (SG) Smith shaves all and only those who don't shave themselves. Question: Does Smith shave himself? If he does, then he doesn't, since he shaves only those who don't shave themselves. If doesn't, then he does, since he shaves all those who don't shave themselves.

OG is no more about God than SG is about barbers. Both paradoxes expose logical problems relating to the use of ***quantifiers.***

Although one of the objectives of the philosophy of religion is to figure out whether or not God exists, its efforts to do this quickly veer off course and end up bogging down in logical minutiae that are of no relevance to religion at all. Consider the following argument, which was given by St. Thomas Aquinas (1225–1274):

> (AG_1[413]) Every event has a predecessor. Therefore, there is some event that is the predecessor of every event. In other words, there is a first event. Causes precede their effects. Therefore, that first event, not being preceded by anything, was uncaused. Since all events must be caused, that first event was a miracle and, therefore, the act of a deity. Therefore, there is a God.[414]

The first sentence of AG_1, which we'll refer to as (a), is ambiguous.[415] It could mean either:

(a_1) Given any event x, there is some event y, such that y precedes x

or

(a_2) there is some event x such that, for any event y, x precedes y.

(a_1) says that any given event has a predecessor. (a_2) says that there is some event that is the predecessor of *every* event.

(a_1) does *not* entail that

(b) there is some event that is the predecessor of every event.

Any given integer has a predecessor. Given only that any given integer has a predecessor, it doesn't follow that some one integer precedes every other—nor is it the case. Similarly, given only that any given event has a predecessor, it doesn't follow that some one event precedes every other.

So (a) fails if (a) is taken to have (a_1) for its meaning, and AG_1 must therefore be assumed to have (a_2) for its meaning. But (a_2) and (b) are the very same statement. So Aquinas has simply *assumed* that there was a first event, and he thus hasn't established it at all.

But supposing, if only for argument's sake, that there was a first event, it doesn't follow, at least not for any obvious reason, that it was an act of God. Why did I hedge with the phrase "at least for any obvious reason"? Because it *may be* that, if there was a first event, there is a God. (I personally do not know this to be false.) It *may* be that atheists cannot, without being guilty of inconsistency, believe in a first event (such as the Big Bang). But even if this is true, it cannot be taken for granted; it must be demonstrated. And Aquinas hasn't demonstrated it.[416]

Aquinas gave another, similarly flawed argument for God's existence:

> (AG_2) There can't be comparatives unless there are absolutes. I'm wealthier than you if I'm more like some absolutely wealthy entity than you are. I'm better than you if I'm more like some absolutely good entity than you are. Some things are better than others. Therefore, some things are more like some absolutely good thing than others. Therefore, there is some absolutely good thing. In other words, there is a God.

AG_2 is a failure. If Smith's complexion is better than Brown's, then

(i) Smith's complexion is more like that of a person with a perfect complexion than Brown's.

But what follows is not that

(ii) somebody has a perfect complexion, and Smith's complexion is more like *that person's* complexion than Brown's,

What actually follows from (i) is that:

(iii) were there to be a person with a perfect complexion, then, other things being equal, Smith's complexion would be more like that person's complexion than Brown's.

Similarly, given only that:

(#) Smith's condition is more like of that some perfect entity's than Brown's,

what follows isn't that

(##) there exists some maximally good entity whose condition Smith's resembles more than Brown's

and is instead that:

(###) were there to be some maximally good entity, then, other things being equal, Smith's condition would be more like that entity's condition than Brown's.

(iii) doesn't entail that anyone's complexion is perfect, and (###) doesn't entail that anything's condition is perfect. Contrary to what Aquinas believes, there don't have to be absolutes for there to be comparatives.

Although (#) is ambiguous, it isn't ambiguous in the same way as:

(JB) Jerry is at the bank.

JB is ambiguous because one of its constituents has two different meanings. In other words, it's ***lexically ambiguous.*** By contrast, (#) is ambiguous *not* because it contains an ambiguous constituent, but because there are two different ways of parsing it. The same is true of (a). Thus, (#) and (a) are ***syntactically ambiguous.*** An expression is syntactically ambiguous iff it can be ***parsed*** in more than one way.[417] (An expression can be both syntactically and lexically ambiguous (for example, "some banks are more like some maximally beautiful bank than others"). Pre-modern philosophers, such as Aquinas, were only dimly aware of the phenomenon of syntactic ambiguity. Hence the problems with AG_2.

Let's take stock. In the course of studying these two arguments, we've learned a lot about sentencestructure and about argumentation in general. But we didn't learn anything about God or religion.

"But we learned that Aquinas failed to prove God's existence." Yes. But his arguments don't fail in a way that gives us any indication as to whether God does or does not exist.

1.1 The futility of trying to provide rational arguments for or against religious belief

Religion is not science or logic, and to the extent that one's beliefs are based on rational consideration of the relevant evidence, one is a person of science, not of religion.

A story will clarify this. Smith hasn't a religious bone in his body. But he's an open thinker and doesn't want to assume, without first exercising due diligence, either that God exists or that God does not exist. On Monday, Smith uncovers what he thinks is evidence of God's existence. But on Tuesday, he finds out that this evidence was manufactured by a religious person and is therefore bogus. Smith doesn't conclude that God *doesn't* exist. Being a scientifically minded person, he rightly concludes only that he doesn't yet have reason to believe otherwise. On Wednesday, he uncovers more apparent evidence of God's existence. Since he finds it convincing, he

begins to warm to the idea that God exists. But, being a rigorous and cautious thinker, he makes sure that his zeal for that hypothesis never outruns the evidence he has for it. And if new evidence turns up, he'll modify his views accordingly.

Smith's belief in God goes only as far as the available evidence permits. For this reason, even if Smith does finally arrive at the belief that God exists, he isn't religious. As long as his belief is contingent on its consistency with future findings, he's prepared to reject it, which is the mark of a scientific, not a religious, person.

Of course, one can be both religious and scientific. Newton was very religious. But to the extent that one's attitude towards a position is characterized by the just-described attitude of deference to future findings, one is not a religious person. Had Newton's beliefs about God been as evidence-based as his beliefs about physics, he would not have been a religious person. A person who believes in God isn't religious *if* that person's acceptance of that position is contingent on its being consistent with the data.

This is why it's silly to try to undermine religious belief by arguing that it doesn't square with the facts. What distinguishes religious from scientific belief is precisely that the former, unlike the latter, isn't subject to one's views as to what inferences the available data warrants.

Religious people know this. That's why they are unmoved by anti-religious arguments. And they're right to be unmoved. To the extent that it's a *religious* attitude they wish to hold onto, it would be incoherent of them to regard recalcitrant data as giving them a reason to jettison that attitude, since a religious attitude is precisely one that isn't marked by the subservience to data and logic characteristic of the scientific view.[418]

1.2 My biases

I am not a religious person and I don't believe in God. (Or so I say—but see Section 1.4.) I say this *not* because I have any special insight into these matters. I do not. My views on these matters are no more authoritative than anyone else's. I am stating these facts about myself only so that the reader can know that, because of them, what I say about religion is biased in favor of an anti-religious perspective.

That said, this chapter won't target religious belief. I will put forth some arguments that might appear to have anti-religious consequences. But, though I obviously accept those arguments, what they prove, in my opinion, doesn't really have anything to do with religion. It has to do with certain attempts to doctrinalize religious sentiment—that is, with attempts to identify religious sentiment with acceptance of well defined *contentions* as to how the world is. Since, as we saw in Section 1.1, religious sentiment necessarily outstrips any logical or evidential reasons one has for accepting any given contention, the logical and evidential demerits of such doctrinalizations are irrelevant to whether the religious outlook is, when evaluated on its own terms, a viable one. So far as one accepts the yardsticks relative to which such doctrinalizations fall short, one isn't religious. After all, the religious mentality is a rejection of those yardsticks. (To be sure, there are religious people who are utterly rational and who are superlative scientists, businessmen, and so on. But to the limited extent to which a given person's mental life is to be understood in terms of his or His having a religious outlook, that individual's psychological situation is not to be understood in terms of his or His ratiocinative abilities, however active those abilities may be outside of religious contexts.) So in attempting to invalidate religion by showing that it fails relative to those yardsticks, one is guilty of "begging the question"—that is, of assuming the truth of the very thing one is trying to prove. Further, given that there is at least one component of religious sentiment to which any given doctrinalization of it *ipso facto* fails to do justice, the *contentions* that are associated with acceptance of a given religion are really just foils of some kind—that is, they are stand-ins for some attitude or conviction that, for some reason, isn't making an appearance *in propria persona.* Instead of asking "how can I disprove the contentions put forth by religious folk?", those who aspire to undermine religious sentiment should ask, "what do such contentions *represent*? What are the sentiments underlying acceptance them?"

To be sure, the integrity of a given religion may depend *up to a point* on the integrity of the doctrines associated with it. Acceptance of a religion typically involves acceptance of well-defined contentions (concerning morality, the structure of the universe, etc.). That said, those who attack religion on scientific or logical grounds—on the grounds that what Darwin or

Einstein said is inconsistent with this or that religious contention— are misguidedly trying to understand religious sentiment in *epistemological* terms, when in fact it is to be understood in psychoanalytic terms. And this, I believe, is part of the reason why the philosophy of religion tends to be so sterile.

1.3 The futility of attempts to liberalize religious belief

A friend of mine who says that He believes in God recently told me that, in His view, "God is love." When I asked His to explain what He meant, He said that, if there is love in the world, there is, for that very reason, a God. God isn't some hyper-person, He explained. He or He or It is nothing other than the fact that people love other people.

This is all very well. But it isn't the view of somebody who believes in God. Freud was an atheist, and he certainly believed that people sometimes love other people. In fact, he said that "libido" is one of the basic forces of human psychology.

Many philosophers have advocated similar positions, saying that God is nothing other than logic, or truth, or reason. I admire the rationalist sensibility underlying these views. But they're not the views of a religious person, since there is no shortage of dedicated atheists who believe in logic—that is, who believe that some statements entail some other statements.

I don't know whether there is a God. I don't even know what it means to say that there is a God. But I know what it doesn't mean. It doesn't mean that some inferences go through; it *doesn't* mean that some people love some people. So God isn't truth, logic, or love—even though, if God exists, that entity, whatever it is, may well be maximally logical, perceptive, and caring.

Attempts to identify God with reason or truth or love are dishonest and incoherent, and so is every other attempt to show that religious sentiment is compatible with 21st century intellectualism. People should cop to what they really believe. If your belief in God extends no further than your belief in logic or libido, then you do not believe in God.

1.3.1 "Never explain, never complain"

Given that religious belief is of its very nature free of the tentativeness and evidence-dependence of scientific belief, religious people have no obligation to defend their views.

Many religious people *have* done so, of course. Many religious people are also scientists, logicians, and philosophers, who, like all other members of those professions, enjoy finding support for their positions. Indeed, some very great philosophers and scientists (e.g., Kant, Newton) have been devoutly religious, and some (e.g., Augustine, Aquinas) have themselves been important religious figures. But *to the extent that one is religious and that one's views are religious views,* one is not only not required to defend one's views, but is actually required to hold that they *need* no defense. This is because a view that is only as good as the logic and evidence underlying it is *ipso facto* a non-religious view.

Rockefeller said: "never explain, never complain." In defending a view that one holds, one is conceding that it needs a defense and therefore that it is to be presumed false until proven correct. When one complains, one is saying that one is powerless, and therefore unworthy, since it is by acting, not whining, that the powerful express themselves.

Obviously there are contexts where Rockefeller's dictum is inappropriate. A doctor or lawyer should surely explain, and a person who is abused should surely complain. But that dictum does hold where religious belief is concerned. Defending a religious view is the same as explaining why one holds it, and explaining why one holds it is almost the same as apologizing for it. So, if you are a religious person, you are entitled to explain your beliefs, if you wish to do so. But know that, in doing so, you are conceding that they are only as good as the arguments for them. And in making that concession, you are saying, however unwittingly, that there but for the grace of the mercurial and unpredictable winds of evidential change go they.

1.4 A conjecture

In his book *Compulsion and Doubt,* Wilhelm Stekel (1868–1940), a psychoanalyst with a gargantuan intellect, said that many self-described atheists are actually crypto-believers. I suspect that, when some philosophers speak of the logical or ethical *standards* to which judgments and acts ought to conform, they are at some level speaking about God. Acceptance of such

rarefied and depersonalized arbiters and right and wrong is sometimes, I would conjecture, a distorted and oblique expression of a belief in God.

1.5 Is there a relationship between the degree to which one is religious and the extent to which one is intelligent?

No. Some of the most intelligent people I know, or know of, are religious and some of the least intelligent are non-religious. (I've known many people who were decidedly unintelligent who, in an effort to convince themselves and others of their intelligence, became rabidly anti-religion.) One of the most brilliant philosophers alive is Michael Dummett. Even though he's devoutly religious, you wouldn't know it from his work. I myself didn't know it until after I'd read most of his works; and I found it out, not by reading one of his books, but by reading an interview with him.

The reason why people of such intelligence may believe in religious doctrines that are insufficiently supported by facts and logic is that religious belief isn't about facts or logic. It's independent of them and isn't answerable to them. The part of the mind that mediates religious sentiment is only in very imperfect communication with the parts that mediate rational thought; and religious sentiment being what it is, it is right that this is so, at least after a fashion.

1.6 The one thing that philosophy has to say about religion

There is only one qualification to my previous statement that philosophy has nothing to teach us about religion. Plato put forth a very clever argument to the effect that, so far as there are moral truths, their basis lies in the structures of properties, and *not* in God's decrees. That argument is known as the "Euthyphro argument," since it was first put forth in Plato's dialogue the *Euthyphro.* I consider that argument to be cogent.

Anthony Flew, a contemporary philosopher, said that there is an easy and reliable way to determine whether a person has an aptitude for philosophy: He does if He understands the Euthyphro Argument; He doesn't if He doesn't. Having had thousands of students, and having gone over the

Euthyphro argument many times in class, I agree with Flew.

There is an easy and reliable way to determine whether a person is *religious*. (But this test only works for people who have studied the Euthyphro argument and who also understand it.) If He *agrees* with the conclusion of the Euthyphro argument, He's not religious; and if He disagrees with it, He is.[419] Let us now discuss that argument.

2.0 The Euthyphro dilemma

Some acts are good and others are bad. Giving money to the needy is good and killing babies is bad. But what *makes* good acts good, and what *makes* bad acts bad? Hise is one answer:

> (G) If an act is good, that is because God likes it. In other words, God's liking it is what *causes* it to be good. If an act is bad, God's disliking it is what *causes* it to be bad. Until God likes or dislikes an act, it remains neutral.

Basically, to be good is to be liked by God, and to be bad is to be disliked by God.

Of course, G assumes that God exists. For argument's sake, let's grant that assumption.

Is G correct?

No. What if God liked rape? Would that make it good? No. Rape would be bad even if God liked it, and God would be bad for liking it.

What if God disliked acts of kindness? Would that make such acts bad? No. They would still be good, and God would be bad for disliking them.

So G is false.

But one might object to this argument:

(CG) There's no way that God would like rape and no way he wouldn't like acts of kindness. You see, God is maximally good; and it isn't even a theoretical possibility that something maximally good should like rape or dislike acts of kindness. Your argument therefore fails, since it presupposes that God could like rape and dislike acts of kindness.

CG doesn't work. Let's suppose, if only for argument's sake, that God is maximally good and therefore dislikes rape and likes acts of kindness. *Why* does God dislike rape? Does he *just* dislike it? Is his dislike of it without any foundation?

No. Surely God has a good reason for disliking rape. But what could that reason be? The answer is clear: rape is bad; rape is wrong. So God's reason for disliking rape is that it is bad and wrong. He sees the fact that rape is wrong; and, on the basis of that knowledge, he judges that rape is wrong.

Rape is wrong. God sees that. And God therefore dislikes rape.

So the fact that rape is wrong is what makes God dislike it. Therefore, it isn't God's dislike of rape that makes it wrong.

For exactly similar reasons, the fact that kind acts are morally right is what makes God like them. Therefore, it isn't God's approval of such acts that makes them right.

Conclusion: Morality is independent of God's views. If, as many of us believe, God is morally perfect and intellectually infallible, then God will like all morally good acts and dislike all morally bad ones. But it isn't God's *liking* an act that makes it good or bad. That act is good or bad all by itself.

Question: Could God *make* rape to be morally good? No. Even God cannot make rape be good. But can't God do anything? So can't he make rape morally good? No, and the reasons are similar to those that make God unable to make a square circle. For an elucidation and defense of this point, see Section 4.0.

2.1 Why, even if correct, the Divine Command Theory is irrelevant to what even maximally religious people believe

Many a sociopathic ghoul has claimed that "God" ordered him to kill. We don't believe such people. But why not? (In this context, it will be assumed for argument's sake that we believe there to be a God and also that we believe God to be morally good.)

One reason is that what the ghoul in question is claiming is inconsistent with our views about morality. "God would have to be evil to tell Hitler to do

those things," we believe. "Therefore, contrary to what Hitler claimed, God didn't tell Hitler to do those things."

Of course, there are often *other* reasons to distrust such people. There is independent evidence of mental illness or of political ambition or of some other factor that would explain why, even though God did not speak to him, the evil-doer in question would say otherwise.

That said, if some otherwise utterly normal person were to tell you that God told him to blow up a school bus, you would, for that very reason, question the credibility of his claim. This shows that *even if* God speaks to certain individuals, it isn't because of what He told them that we believe that rape, theft, and so on, are wrong. If I tell people that God personally told me that x is wrong, I'll be seen as a fraud or a madman by those who think that x is good and who therefore think that God would never condemn it, and I'll have no effect on those who already think that x is wrong.

If somebody claims that God said X, it is only by considering the merits of X itself that we can determine whether that claim is legitimate or not. This means that our decision to accept or reject X is *not* based on what God says. So even if God believes that X is good and directly tells certain people this, that isn't why we believe that X is good. Our views as to what is right and wrong don't have anything to do with what God believes. So even if it's correct, the Divine Command Theory can do nothing in the way of legitimating existing religious practices or beliefs.

3.0 The problem of evil

Many people hold that there is a God and that He is all-powerful, all-good, and all-knowing. For argument's sake, let us suppose that this is true. In that case, we are stuck with a problem. Many bad things happen in this world (e.g., murder, disease, physical and emotional pain). But if God is all-knowing, all-powerful, and all-good, then whenever something bad is going to happen, He'll know it (since He's all-knowing); He'll also want to prevent it (since He's all good); finally, He'll *be able* to prevent it (since He's all-powerful). Nonetheless, there obviously *are* many bad things in this world. This suggests that there is no God or that, if there is a God, He isn't all-powerful, all-good, and all-knowing.

Here are some views that people adopt in order to make sense of the information just put forth:

(a) There is no God.

(b) There is a God, but He's not all-powerful, all-good, and all-knowing.

(c) There is no badness—there is only the *illusion* of badness. If we were to know God's master- plan, we'd see that it's okay that people have their arms and legs blown off.

But there's a fourth view, and it is the view that, from a philosophical standpoint, is the most interesting:

(d) God exists, and He *is* all-powerful, all-good, and all-knowing. Furthermore, there is much evil in the world. *And these statements are consistent with one another.* There is nothing absurd in believing in the existence of badness *and* in the existence of all-powerful, all-good, and all-knowing God.

But we saw earlier that (d) *appears* to be incoherent. If there is a God, and He's all-powerful, all-good, and all-knowing, then He'll prevent bad things from happening. But bad things do happen. So it *seems* to follow that either God isn't all-powerful, all-good, and all-knowing *or* He simply doesn't exist.

But many fine thinkers hold that (d) is *not* incoherent. They hold, in other words, that (d) makes perfectly good sense. In their view, it is perfectly reasonable to hold that, despite all the badness in this world, there exists an all-powerful, all-good, and all-knowing God.

Attempts to show that (d) is coherent are known as *theodicies.* Following are the two most plausible theodicies.

3.1 Theodicy #1—The Free-Will Defense

Because He is good, God wants the world to be a good place. In fact, He wants it to be as good as possible. A world where nothing has free will wouldn't be a very good world, and it wouldn't be *as* good as one where there were creatures with free will. A world of robots who do no wrong isn't as good as a world of free creatures that do some wrong. (Would you

> rather have a son or daughter who is a robot who never does anything wrong but doesn't have a soul? Or would you rather have a son or daughter who has a soul but sometimes parties too hard?) Not wanting us to be soulless robots, God gave us free will. But now that we have free will, we're free to commit bad acts—to steal, kill, and so on. And that's why, despite God's goodness, there's badness in the world.

There are three problems there.

3.1.1 First problem: Theodicy #1 isn't compatible with the compatibilist conception of freedom

In Chapter 15, we saw reason to believe that, not only is determinism compatible with freedom but is actually a prerequisite for it. If this is the case, then, even if we are free, God is still responsible for the badness that we, in our freedom, inject into the world.

But even if we set aside this problem, there are two other grave problems with Theodicy #1, as we'll now see.

3.1.2 Second problem: There's too much evil for Theodicy #1 to be correct

Couldn't God allow us to have free will *in general* but intervene in cases where somebody was going to commit particularly horrible acts or was going to commit horrors on a mass scale?

Here's how advocates of Theodicy #1 deal with this:

(AF[420]) Suppose that God thwarted every would-be Hitler and Stalin. Suppose, in other words, that God always saved us from ourselves at the moment of truth. In that case, we'd be in a state of perpetual infancy—we'd never have to learn to manage our own affairs. God doesn't abridge the free will of Hitler and Stalin because, were he to do so, he'd be running the show, instead of giving us a chance to demonstrate and develop our characters by running the show for ourselves.

Whatever its merits, there is a very obvious problem with AF, which we'll

now discuss.

3.1.3 Third problem: not all badness is our doing

The third problem is not all badness results from *human* activity or, therefore, from the exercise of free will. There are diseases, floods, earthquakes, etc. The Free-Will defense has nothing to say about these, since it isn't our free will that causes these bad things to happen. So couldn't God spare us these horrors? And wouldn't he do so if he really were all-good, all-powerful, and all-knowing? The response to this question lies in:

3.2 Theodicy #2—the no pain/no gain defense

Suppose that every time a disaster was about to strike, God stepped in. In that case, life would be a big cakewalk, and we, as a species, would never develop any character. We'd have it too easy. Our ingenuity and fortitude would never be put to the test. We'd be deprived of a chance to prove and *im*prove ourselves, which would strip life of much of its meaning.

If life were easy, it wouldn't be a challenge; if it weren't a challenge, it wouldn't be meaningful.

A certain amount of strife enriches life. But too much impoverishes it. There's too much strife for this theodicy to work. That's my position, anyway. Others disagree.

4.0 Can God do the impossible?

Even if, as we'll assume in this section, God exists and is omnipotent, He is unable to make 1 + 1 = 3.

For something to be impossible is not for it to be very hard to do: it is for the very idea of doing that thing to be an *incoherent* one. It is, in other words, for there to be no such thing as doing that thing. Something is impossible just in case nothing would *count* as doing it—just in case, in other words, no meaningful statement that would describe the occurrence of that thing.

Nobody can read a thousand-page book in two seconds. But there is some sense in which it would be possible, albeit very hard, to do so. The reason it is possible to read such a large book in so little time is that the idea

of doing so is not *incoherent.* We know what it is to read a long book in (e.g.) a week. On that basis, we know what it would be to read such a book in a day or an hour or even a second. So, even though nobody can read a thousand-page book in two seconds, the *idea* of somebody's doing so is not an incoherent one; and that is why somebody's doing so is possible. (Of course, to say that it is possible is not to say that it is *practically* possible. It is only to say that it is *logically* possible.)

By contrast, it seems impossible—not just practically, but logically—that something should preexist itself. It seems impossible, in other words, that something should come into existence before it comes into existence or, if it is an event, that it should occur before it occurs. What is the difference between somebody's reading a thousand-page book in two seconds and an event's occurring before it occurs? Why is the one possible and the other impossible?

4.1 Impossibilities are expressed by self-undermining statements and, therefore, by non-statements

The answer is not hard to find. The statement:

(PE) X preexists itself

denies the very thing that it affirms. It says (1) that, for some time t, X began to exist at t and (2) that, for some earlier time t*, X began to exist at t*. In saying (2), PE is saying that X did *not,* in fact, begin to exist at t. So PE says that X begins to exist at t, and it then says that X does *not* begin to exist at t. PE therefore says that the very thing (or, in any case, one of the very things) that it says is false.

Thus, the reason why it is impossible that X should preexist itself is that the very idea of X's doing so is self-defeating. To think that X preexists itself is to think (A) that X begins to exist at time t and also (B) that (A) is false. Since such an idea is self-contradictory, it fails to say *anything* coherent about reality.

What we just said about the statement "God preexisted himself " is

true of any other statement that describes an impossibility. For example, nothing can be a square and a circle at the same time. Therefore:

(SC) "X is a square circle"

describes an impossibility.

But can *God* do the impossible? Can *God* create a square circle? I say "no." And I also say that, even though God cannot create a square circle, this does not show that His powers are in any way limited.

Let us look at SC more closely. A square is by definition something that has four straight sides. A circle is by definition something that has no straight sides. So (SC) makes two statements: first, that X has four straight sides; second, that it has *no* straight sides (and therefore doesn't have four straight sides). So SC says that the very statement (or, at any rate, one of the very statements) that it affirms is false. So it affirms a statement and then withdraws it, thus failing to make *any* statement.

Thus, when we say that God cannot make SC be true—when we say, in other words, that God cannot create a square circle—there is no coherent statement that we are saying that God cannot make true; and there is, therefore, *nothing* that we are saying that God cannot do. So in saying that God cannot create a square circle, there is no coherent statement that we are negating, and we are therefore not managing to ascribe any genuine limitation to God's powers.

4.2 The confusion underlying the traditional conception of what it is to be impossible

We pre-theoretically tend to think of impossibility as a property of *things.* Consider the following statements:

(SC) Square circles are impossible—so are perpetual motion machines, self-causing events, and even primes greater than two.

SC is a true statement. But, at first glance, SC seems to be self-defeating. For it seems to be saying that there exist certain things—namely, square circles, perpetual motion machines, and even primes greater than two— that cannot,

and therefore do not, exist. In other words, SC seems to be saying that there exist things that don't exist. But if that is what SC is saying, then it is not only false, but incoherent.

But it is not incoherent. In fact, it is true. How can we make sense of all this?

The key Hise is to distinguish between *objects* and *statements.* SC doesn't make a statement about *objects:* it makes a statement about *statements.* It says that all statements of the form *X is a square circle* (and *X is a perpetual motion machine* and *X is an even prime greater than two*) are false.[421] So SC doesn't say of any object that it is, or is not, a perpetual motion machine. It couldn't possibly say this, since there are no perpetual motions

and there are therefore no perpetual motion machines of which to say that they are impossible (or, for that matter, of which to say *anything*). Rather than making a statement about non-existent perpetual motion machines—and thus about things that don't exist to have statements made about them (or, for that matter, to have anything else done to them)—SC makes a statement about very much existing statements. More specifically, it says that every statement having the form *X is a perpetual motion machine* is untrue.

In general, impossibility is not a property of *things.* After all, if something is impossible, then it cannot, and therefore does not, exist. And if it does not (let alone cannot) exist, then there exists no thing to have the property of being impossible. Impossibility is a property, not of things, but of statements. And, as we saw in the previous section, a statement is impossible if it is *incoherent*—if, in other words, it contradicts itself.

4.3 Explicit versus implicit self-contradiction

We must bear in mind that most incoherent statements contradict themselves *implicitly,* not explicitly. The statement:

(KB) X is knowledge, but it isn't true belief

is incoherent. But it does not *overtly* or *explicitly* have the form *P and not-P.* Nonetheless, it *implicitly* has that form. Since any instance of knowledge is necessarily an instance of true belief, KB does, albeit implicitly, say that X

simultaneously is, and is not, true belief. Therefore, KB *does,* albeit implicitly, say *X is true and it is not the case that X is true belief.* So KB does in fact have the form *P and not-P,* that being why it is incoherent.

Let us revisit the topic of God's omnipotence. To say that God cannot do the impossible is therefore *not* to say that there exist certain *things* that God cannot do or objects that He cannot create. There exist no impossible *objects.* If something exists, then it isn't impossible. But there are incoherent *statements.* And when we say that "square circles" are impossible, we are saying that *statements* of the form *X is a square circle* are selfundermining. An incoherent statement is one that does not, ultimately, say *anything* (since it denies, and therefore retracts, what it affirms). So there is nothing that God's "inability," so to speak, to create square circles is an inability to create.

4.4 Incoherence versus meaninglessness

A moment ago, I said that, ultimately, an incoherent statement is one that doesn't say anything. This statement must be understood aright. The statement:

(A) X is a square circle

is obviously, at some level, meaningful; and it obviously does say *something.* An example of a completely meaningless utterance—one that simply fails to say anything at all—would be:

(B) Frog the marathon pool clearly dabble.

It is clear that (A) is incoherent. But it is not clear that (A) is meaningless. Indeed, (A) seems to be far from meaningless. By contrast, (B) seems to be quite meaningless—so meaningless, in fact, that it's not even clear if it can be described as "incoherent." (A) and (B) are obviously not in the same category.

A related point is that, whereas (B) doesn't say anything, (A) very much *does* seem to say something, albeit something defective. And this, of course, is why (A) and (B) are clearly not in the same category, notwithstanding that they are both failed statements. What, then, is the difference between incoherence and meaninglessness? And in what sense, if any, does (A) fail to say anything?

For a statement to be meaningless is for it to have *no* meaning. For a statement to be incoherent is for it to have two (or more) *opposed* meanings. (B) has *no* meaning. It doesn't say anything. By contrast, (A) has *two* meanings. (It says, on the one hand, that X has four sides and, on the other, that X has no sides and therefore less than four sides.) So, in the short term, (A) *does* have a meaning—indeed, it has two of them. But *ultimately* it has no meaning, since those two meanings cancel each other out. So *ultimately* (A) fails to say anything. (Saying two opposed things is a way of failing to say anything—just as taking one step forward and one step backwards is a way of failing to go anywhere, even though it is a way of failing to go anywhere that *involves* going to two places.) Therefore, my previous statement—that incoherent statements *ultimately* fail to say anything— has turned out to be true, and also capable of being reconciled with the fact that it clearly does, at some level, have some kind of meaning (unlike (B)).

5.0 Anselm's version of the ontological argument for God's existence

The so-called "ontological argument" for the existence of God was first put forth explicitly by St. Anselm. There are different versions of this argument. We'll consider three, starting with Anselm's own.

> (AG[422]) Let us define "God" as a "maximally great being." (In this context, a being is "maximally great," if it isn't possible for me, or you or whoever, to conceive of anything greater than it.)With this in mind, consider the following argument. I can conceive of a maximally great being. Thus, *a maximally great being exists in my mind.* For argument's sake, suppose that such a thing existed in my mind, but not also in reality. In that case, it *would* be possible to conceive of something greater than that thing—namely, something existing in reality but otherwise just like that thing. (A real mansion is better than an otherwise identical imaginary mansion. Similarly, a real maximally great thing is better than a merely imaginary one.) It is therefore incoherent to suppose that a maximally great thing exists in my mind but not also in reality. So given that such a thing exists in my mind, it must also exist in reality. Conclusion: a maximally great thing exists. In other words, God exists.

AG is a spurious argument. The problem lies in the italicized statement, which we'll call "MG." Correctly interpreted, MG doesn't say that God exists *in my mind.* If God exists, He is out there, in the world, not in my mind. No idea of mine is God. The only things that exist in my mind are thoughts, feelings, etc. Some of these things are *representations* of things that exist outside of my mind. But the fact that a *representation* of the Sun exists in my mind obviously doesn't entail that the Sun itself exists in my mind. And the statement "the Sun exists in my mind" is a brazen falsehood, unless it's misleading of saying that a Sun-idea so exists. Similarly, the statement that God exists "in my mind" is patently false, unless it's a misleading way of making the irrelevant point that a God-*idea* so exists. So, to the extent that it's true, MG doesn't entail that God exists in any way, shape, or form; and, therefore, no contradiction from the supposition that MG is true but that God doesn't exist "in reality."

Here's a similar line of thought. Correctly interpreted, MG says that I know how a thing must be if it is to be maximally great. In other words, it says that I know what conditions a thing must fulfill in order to be maximally great.

But given only that I have this knowledge, it doesn't follow that there exists anything that has any of the requisite properties. It doesn't follow that there is anything that is all-powerful or all-knowing or all-good or that has any other virtue specific to God. AG goes through only if there exists something that has these properties. The italicized statement must guarantee that there is such a thing; otherwise the argument collapses. It doesn't, and the argument collapses.

5.1 A simpler version of the ontological argument

Step 1. God is a maximally excellent thing.

Explanation: Just as squares are, by definition, four-sided, so God is, by definition, maximally excellent. If we found that object X was deficient in some way—that it was ill-willed or stupid or lazy—then it would, for that very reason, fail to be God.

Step 2: Because he is a maximally excellent thing, God has all excellences, and he has them to the highest degree.

Explanation: Consider some excellence—for example, physical strength. And suppose that X is puny and weak; that is, that X has physical strength only to a small degree. Obviously X isn't God. So God must be maximally strong and, for similar reasons, must be maximally smart, kind, good-looking, etc.

Step 3. Existence is an excellence.

Explanation: There are a lot of ways that one can fail. One can be lazy, stupid, mean, etc. But no failure is as bad as the failure to exist. If you don't exist, you are a nothing, a complete and total zero, and you are therefore a nothing *compared* to anyone—even the biggest loser you know.

Conclusion: God must have existence. God is, by definition, maximally excellent and therefore has every excellence. Since existence is an excellence, it follows that God has existence.

5.1.1 This argument evaluated

No good argument can assume the truth of its conclusion: arguments mustn't "beg the question." But the ontological argument appears to do just this. The problem lies in the first two steps. Step 1 is just another way of saying:

Step 1*: There exists a certain thing, namely God, which is maximally excellent.

And Step 2 is a way of saying:

Step 2*: There exists a certain thing, namely God, which, because he is maximally excellent, has every excellence to the highest possible degree.

But if Steps 1 and 2 are taken in this way, then the argument question-beggingly *assumes* that God exists, and thus fails to prove it.

So, if the Ontological argument is to go through, Steps 1 and 2 must be interpreted in such a way that they don't assume that God exists. Thus interpreted, their meanings are:

Step 1#: *If* God existed, then He would be a most excellent being.

Step 2#: *If* God existed, then God would have every excellence to a maximally high degree.

So if the ontological argument is to avoid begging the question—in other words, if it is to avoid simply *assuming* God's existence—then we must replace Steps 1 and 2 with Steps #1 and #2. So the first three steps must become:

Step 1#. *If* God existed, He would be a most excellent being.

Step 2#: *If* God existed, then God would have every excellence to a maximally high degree.

Step 3. Existence is an excellence.

What follows from these three statements is that:

Conclusion*: *If* God exists, then He has every excellence, including existence.

In other words, if God exists, He exists.

But that statement is completely innocuous. Even atheists grant that *if* God existed, then He *would* exist.

So the ontological argument proves nothing.

5.1.2 The simplest possible version of the ontological argument (SOA)

Failing to exist is a way of being imperfect. God is perfect. Therefore, God doesn't fail to exist.

5.1.3 This argument evaluated

SOA presupposes that a given thing's failing to exist makes it imperfect. SOA therefore presupposes that a given thing can fail to exist. This seems reasonable enough at first. ("Wouldn't I be a lesser person if I failed to exist?") But it's incoherent. If you didn't exist, you wouldn't be a lesser person. There would be no you to begin with; and there would thus be no you to fail to do or be this or that.

For there to be something that failed to exist would be for there to exist something x such that x did not exist. And that, clearly, is not possible.

“But surely there are things that don’t exist,” it will be said. “Square circles don’t exist. Neither does Zeus.”

This position is dealt with at length in Chapters 1 and 8. But it’s easy to say what’s wrong with it.

The sentence:

(SC) “square circles don’t exist”

doesn’t say that

(SC#) there exist objects that are both square and circle; and those objects don’t exist, which would be nonsense. Rather, SC says that

(SC*) any given thing is a non-square if it’s a circle

In other words, pick any given object: if it’s a circle, it isn’t a square. And the sentence

(ZE) “Zeus doesn’t exist”

doesn’t mean that:

(ZE#) there exists is a certain thing, viz. Zeus, that doesn’t exist.

ZE makes the innocuous claim that given any object x, x doesn’t have the property of being a unique,

lightning-bolt hurling god [etc.].

Thus, SC and ZE attribute properties to properties. They don’t attribute non-existence to objects.

And:

(CS) square circles do exist

and

(EZ) Zeus exists

are to be understood similarly. There isn’t some square circle x such that CS is to the effect that:

(MCS[423]) x exists.

Rather, CS is to the effect that:

(MCS_1) it isn't true that everything that is a square isn't also a circle

or, alternatively,

(MCS_2) the set of circles overlaps with that of squares.

And there isn't some god x such that x hurls lightning bolts (etc.) such that EZ is to the effect that:

(MEZ) x exists.

Rather EZ is to the effect that:

(MEZ) it isn't true of any given thing x that x doesn't uniquely have the property of being something that lives on Mt. Olympus and hurls lightning bolts [etc.].

It isn't even *possible* to attribute non-existence to a non-existent object. One's doing would so would involve there existing some non-existent object to which one attributed non-existence. But that isn't possible. Nor is it possible to attribute existence to a non-existent object. One's doing so would involve there existing some non-existent object to which one attributed existence. But that isn't possible.[424]

Thus, nothing to which existence or non-existence is attributed is non-existent. This is a pretty innocuous claim, as it's identical with the obvious truth that there is no object x such that existence or non-existence is attributed to x and such that x doesn't exist.

It follows that a precondition for a thing's being described as existent or non-existent is its existing. This means that, if P is the proposition one affirms when one describes a given object as existent, there is no negation of P. Since any given proposition can be negated, it follows that, in attributing existence to a given thing, one isn't affirming anything. It is meaningless to say of specific individuals that they exist or don't exist.[425]

Chapter 26

Existentialism

1.0 Existentialism: an attitude more than doctrine

The traditional view is that, when a person comes into the world, there are various standards to which he must conform and various values that he must uphold. No sooner does a person acquire the ability to speak than the weight of centuries of tradition bears down on him, forcing him to do things he doesn't want to do and live in a way he doesn't want to live.

Existentialism says: "To hell with all these preexisting values!" Existentialism is thus an injunction that we choose our own values and, more generally, chart our own course.

At its core, existentialism is therefore less of a doctrine than it is an attitude. A doctrine is a system of interconnected propositions. Quantum physics is a doctrine. So are Freudian psychology, Keynsian economics, and Russell's theory of descriptions. The viability of a doctrine has nothing to do with how anyone feels about anything. Doctrines are right or wrong. They are to be judged according to how well they measure up to the relevant logical or empirical benchmarks. Opinions and feelings have nothing to do with it.

Attitudes cannot be understood in these terms. If you tell me that you like being at the beach, or that you hate picnics, I cannot say: "you're wrong" or "you're right." By contrast, if you tell me that you think that the U.S. economy would improve if a Marxist government were to take over, I *can* say "you're wrong" (or "you're right"). So to the extent that it is an *attitude,* existentialism cannot be judged to be wrong.

My own view is that, taken as an attitude, there is much to be said for existentialism. If you come into the world believing that values are to be chosen, as opposed to passively accepted, you are probably more likely than you would otherwise be to reflect on what your values should be and also, consequently, on what the merits (or demerits) are of traditionally accepted values. And such reflectiveness is positive in two ways. First, if you *choose* your values, instead of simply accepting the values of others, you are (all other things being equal) more likely to live according to values that work for

you psychologically. Second, even if you choose the wrong values, the mere fact that you *chose* your values, instead of passively and robotically accepting the values of others, would seem to be an affirmation of your rationality and freedom and, therefore, of your humanness.

1.1 "Don't believe the hype"—the message of existentialism

There is actually a third reason why it is (or, at least, may sometimes be) better to choose one's values than to accept the values of others. A number of values seem to involve a diminishment of self:

"Be a team player—don't do what *you* want."

"Work for the family business—don't do what *you* want."

"Take care of your ailing mother—don't do what *you* want."

"Believe what you're told to believe in—don't have your *own* beliefs (you're not qualified, you're not smart enough, you don't have the right educational background, and it's not your place to have your own views even if you are qualified: who do you think you are!)."

It would not be hard to think of a case where flourishing and following conventional values failed to coalesce. So and so wants to be a writer. Supposedly wanting him to have a respectable life, so and so's parents (or teachers or priest . . .) tell him to take the safe path ("be a lawyer," "sell insurance"). So and

so becomes Tolstoy (or Shakespeare or whoever). Had so and so followed the conventional path, so and so would have lived a life of "quiet desperation," to use Theroux's expression. (Indeed, Theroux used that very expression to make a point not unlike ours, and Theroux is sometimes seen as one of the first existentialists.)

Why do traditional values so often demand a reduction of the self? One possible answer is that the values that are endorsed by society are those that are good for society, and therefore aren't necessarily those that are good for the individual. As Freud emphasized in *Civilization and its Discontents,* society is possible only if the aspirations of individuals are abridged: a precondition for civilization is the abridgment of human freedoms and,

consequently, of gratifications. A related point—one that presupposes a rather dark conception of human nature and may therefore meet with some resistance—is that the values that society asks the individual to accept embody much of the antagonism that people have for one another. People's lives often aren't what they want to them to be, and the values that people impose on one another (e.g., "be a team player," "it's not about what you want; it's about what's right for the family") often express their desire to prevent others from superseding them.

If you see values as yours to accept or reject, you are less likely to be duped by moralistic propaganda into reducing yourself, and your life is more likely to follow the path that it *should* take, as opposed to the one that others want it to take. In any case, that is the idea—and I'm pretty sure that there is at least some truth to it.

2.0 Existentialism doctrinalized

So far so good. Given only what we've said, existentialism seems to be viable. But several problems arise. Although existentialism is not *merely* a doctrine—although it is fundamentally an attitude, as opposed to a belief-system—existentialist philosophers have spent a great deal of time producing doctrines (belief-systems) that validate the attitude in question. Even though that attitude may be salutary, it doesn't follow that those doctrines are correct or even coherent. Since we are concerned with existentialist *philosophy*—that is, with the efforts of existentialist philosophers to provide a rational foundation for the attitude that we've been discussing—we have to assess whether those doctrines are any good.

But before we consider the viability of those doctrines, we must note there is also a problem (actually, a couple of problems) with the existentialist attitude itself. The values that other people tell us to accept may not always be the right ones. But it doesn't follow that there are *no* values, and it doesn't follow that one is not, by the mere fact of being human (and therefore sentient and rational and intelligent), answerable to certain ethical norms. Sometimes when people reject conventional values, it is because they feel that they must do so to uphold *other* values. In cop movies, it often happens that the hero decides to reject one value ("be a team player—do what your

lieutenant tells you to do . . . not what *you* happen to think would serve the interests of justice") in order to uphold another, presumably more important value ("find the actual criminal: serve the interests of justice—forget about the politics internal to the police department"). It would be hard to think of a single case where a fruitful rebellion against conventional values did not involve an acceptance of what (at least in the mind of the rebel in question) was a deeper and more legitimate set of values. So while there are obviously cases where one oughtn't accept the values that others want one to accept, it doesn't follow that there are *no* values that one ought to accept.

We said earlier that attitudes are not themselves true or false. (Or so, for argument's sake, we'll momentarily assume, in keeping with the conventional wisdom. See Chapter Section 3.0 for an alternative viewpoint, which is outlined in Section 4.0 of the present chapter.) But, despite this fact, attitudes often involve *presuppositions* that are true or false. Suppose that you hate Tom. Your hatred of Tom is not itself true or false. But your reasons for hating him presumably lie in a belief that you have—you hate him (let us suppose) because you believe that he stole your bicycle. And that belief is true or false—even though the same is not (at least not obviously) true of your hatred itself. In general, our emotional attitudes—though (arguably) not true or false in and of themselves—are *grounded* in beliefs about the world, and those beliefs *are* true or false.

To take another example, you really love Tom. Your love of Tom is not, in and of itself, true or false. But your affection for Tom (probably) has a basis in beliefs that *are* true or false. You think that he is a great writer and also that he is a decent human being. Should you discover that he actually plagiarized his award-winning

book, and that his kindness is just an act, you will stop liking him. So, even though your affection for Tom is not true or false in and of itself, it is rooted in beliefs that are true or false. In general, attitudes—though not true or false themselves—are rooted in beliefs, and those beliefs *are* true or false.

Let us reconsider existentialism in light of these points. Considered as an attitude, existentialism is not true or false. But it seems to presuppose a thesis that is true or false, namely:

(EX) there is no set of values that one *must* accept.

But (EX) is ambiguous, as it could mean either

> (A) there is no set of values that, *morally* speaking, one must accept (i.e., there is no set of values that one must comply with *in order to be a good person*)

or

> (B) there is no set of values that, *psychologically* speaking, one must accept. In other words, acceptance of any one value-system is no more, and no less, likely to bring one happiness (or, more generally, fulfillment) than acceptance of any other value-system.
>
> Let us evaluate these theses, starting with (A).

3.0 Does it make sense to speak of moral values that one is *morally free* to reject?

Obviously people do accept certain values, and reject others. But if there are moral values, it's unclear how they could be optional. Suppose that it really is wrong to commit murder. In that case, (A) seems to be saying that it isn't morally incumbent on one to accept the moral prohibition against murdering. But that is the same as saying that, morally, one *can* murder. So if (A) is right, then *there is no moral prohibition* against murdering. Given any moral law, an exactly similar argument shows that, if (A) is correct, then that moral law doesn't really exist. So (A) really amounts to: there is no such thing as morality. But in that case, supposing that (A) is the correct disambiguation of the conceit underlying existentialism, existentialism is only as plausible as the position that there are no moral truths.

But *are* there in fact no moral truths? Isn't it wrong to kill babies, to steal from people in comas, and so on? A popular answer is:

> (MR) Those things are wrong—but only *relative* to our culture (or species). Morality is culturally relative or species relative. In any case, it's relative in some way or other: there are no *universally valid* moral truths. Granted, we *feel* that it is just plain wrong to kill babies. But that doesn't mean anything. A lot of people feel a lot of ways about a lot of things. Not all of those feelings are right. Some people like

> chocolate, others like vanilla. Is it more right to like chocolate than vanilla? No. Some cultures (or species) dislike murder, others like it. Is it more right to dislike murder than to like it? No. Statements like "murder is just plain wrong" are expressions of *feelings,* not of objective fact.

But presumably we *feel* that it's wrong to kill babies because we *believe* that it is wrong to do so. MR presupposes that feelings cannot possibly have any objective basis. But this isn't so. In any case, in the final analysis, nobody really, in their heart of hearts, believes that to be so. Our feelings seem to track our feelings about rightness and wrongness. We tend to be extremely angry at people who we believe to have wronged us. This is not to say that our feelings are always appropriate. But given only that our feelings are often misguided, it doesn't follow that they are categorically without an objective basis. (Our perceptions aren't always completely accurate. But that doesn't mean that they are *always* completely wrong. The same goes for our beliefs. So why suppose that feelings are *always* wrong, given only that they sometimes are?) In any case, it is clear that (A) is, at the very least, extremely questionable, the reason being that there is a strong pre-theoretic presumption to the effect that there are objective moral truths.

Before we move on, we should note that the word "value" is ambiguous, as it has (at least) three meanings.

3.1 Meaning #1

A "value" can be something that is *in fact* valued, or it can be something that one *ought* to value. In the sentence:

(LH) a life of honest toil has value, whereas a life of deception and chicanery does not,

the word "value" doesn't mean "something that *people* believe to be valuable" and it certainly doesn't mean "something that people like." Suppose that everyone alive were a sociopath (a completely amoral person), who not only didn't *want* to live honestly, and wouldn't enjoy doing so, but didn't even believe that such a life would have any value. (LH) would not, at least not on that account, be false in such a world. (In a world of

mathematical morons, nobody would know that 2 + 2 = 4; but the statement "2 + 2 = 4" would not on that account be false. Similarly, in a world of "moral imbeciles" (the 19th-century psychiatric term for sociopath), (LH) wouldn't be false (in any case, it wouldn't be false *in virtue of* the fact that nobody in that world had any desire to live honestly).

3.2 Meaning #2

A "value" can be something that somebody *believes* (perhaps wrongly) to be valuable. People often come to believe that they have been abiding by the wrong values, and must therefore adopt new ones. Tolstoy's novel *The Death of Ivan Ilyich* illustrates this point. Up until the moment right before his death, Ivan Ilyich thinks that what is valuable is being socially respectable, not rocking the boat, and making sure that one's thoughts and deeds meet with general approval. Ilyich is by no means a bad person, at least not in any obvious sense. He doesn't commit crimes. He is a good father and husband. Further, Ilyich is professionally successful, thus enjoying the respect of his colleagues, and he is also wealthy, thus enjoying a materially comfortable lifestyle. By any conventional measure, Ilyich's life is a good one: good from a professional standpoint, good from a moral standpoint, and good from a hedonistic standpoint. Nonetheless, the moment before he dies, it occurs to Ilyich that his entire life was one big lie and that it was, as he puts it, "not the real thing." It is not, I repeat, that Ilyich did anything *bad* (at least not in the conventional sense of that word). He didn't kill anybody; he didn't abuse his wife; he didn't take bribes. It is just that, throughout it all, he wasn't really being himself; he was thinking in terms of how others would see him, in terms of what was "proper." So, right before he dies, he comes to think that what he *should* have valued was being oneself. Ilyich's views as to what is *in fact* valuable thus underwent a dramatic change. His "values" changed in the sense that his *beliefs* as to what is of value changed.

3.3 Meaning #3

A "value" can be something that somebody simply *likes*—but that isn't necessarily valuable and that the person in question doesn't necessarily even *believe* to be valuable. Ted Bundy obviously valued committing heinous

crimes. But did Ted Bundy believe that doing so *had value?* It is not clear how to answer that question.

3.3.1 Commentary on Meaning #3

Some authors hold that one cannot like something without believing it to have objective value. This position is sometimes known as "value-internalism." The value-internalist says that, since Ted Bundy liked committing murder, he must have seen doing so as having value. Other authors hold that someone can like something but regard it as being entirely without value. This position is sometimes known as "value-externalism." According to what the value-externalist says, given only that Ted Bundy *liked* committing murder, it doesn't follow that he saw doing so as having any objective value. I am strongly inclined to agree with the value-*in*ternalist. (In a moment, we will present a value-internalist critique of existentialism—or, more specifically, of B.) I think that people must value the things that they like. But I also believe that a single person can have conflicting values. So the same person can simultaneously regard a certain kind of act as having both positive value and also as having negative value, or no value at all. But my position is by no means universally accepted. We will develop these points later on.

In any case, given that the word "value" is ambiguous between these three different meanings, it follows that (A) is correspondingly ambiguous. Which is obviously something to keep in mind when trying to determine whether (A) is correct or not.

4.0 Are certain values hardwired into us?

Let us now discuss the position that:

> (B) There is no set of values that, *psychologically* speaking, one must accept. In other words, acceptance of any one value system is no more, and no less, likely to bring one happiness or fulfillment than acceptance of any other value system.

A little while ago, we saw why, at least arguably, our emotions embody certain value-judgments and thus presuppose acceptance of a certain system

of values. It's hard to escape the feeling that, if you like something, it's because you think it is good, and that if you hate something, it's because you think it's bad. Bill stole your bicycle, and you're therefore mad at him. Surely, your being mad at Bill reflects your belief that Bill *wronged* you. Bill buys you another bicycle, apologizing in a clearly sincere manner for his act of theft, and you therefore like him again. Surely, your liking Bill again reflects your belief that Bill has behaved justly towards you. In general, it's hard to believe that emotions don't track value-judgments—hard to believe, therefore, that they aren't *identical* with such judgments.

In a moment, we'll discuss some of the objections that are to be made to this. But for now, let us explore the consequences for (B) of supposing this obviously reasonable—though debatable—thesis to be correct.

Supposing that our emotions embody value-judgments, it seems to follow rather straightforwardly that we cannot choose *any* value-system without jeopardizing our own happiness. Consider the story told a moment ago. First you were unhappy with Bill, this being a reflection of your view that, in stealing your bicycle, Bill was acting unjustly towards you. Then you're happy with Bill, this being a reflection of your view that, in apologizing and buying you a new bicycle, Bill has acted justly towards you. It thus seems that your emotions presuppose certain ethical views.

If, in fact, your emotional state—in particular, whether you are happy—is a reflection of your having a certain *moral* outlook (i.e., of your having a certain value system) then you can't change your value system without changing your emotions. Thus, a necessary condition for really accepting a new set of values would be that you re-wire, so to speak, your emotions—that you change how you emotionally react to situations.

In light of these points, let us reconsider (B). Your innate psychological structure obviously puts limits on what you can do and still be happy. Your emotional state—and, in particular, your level of happiness—tracks your views as to the moral fitness of facts about your life, a corollary being that your emotions incorporate a certain value system. In any case, this is what we just argued; and, if it's correct, then that value system is "wired into" your emotional constitution (i.e., you are wired to have certain values) and you thus would be able to adopt a new value system only at the expense of your own emotional well-being. To

sum up, given the

reasonable view that what a person feels is not entirely independent of that person's views as to the moral fitness of facts relating to his or her life, it follows that certain value systems are internal to his/her emotional architecture. In that case, given our emotional architecture, we *cannot* accept just any value system and still be happy.

This doesn't mean that all people have the same values—clearly they don't. But what it means (or at least suggests) is that the *range of admissible* value systems is limited by immutable facts about our psychology (and biology).

4.1 Morality and language compared

Obviously there are different viable languages—in fact, infinitely many different (possible) viable languages. But Chomsky showed that any humanly learnable language satisfies rather narrowly defined formal constraints. And, as Chomsky points out, this strongly suggests that the range of languages that are viable (for humans) is set by innate facts about our psychological-biological structure—that we are "hardwired" to learn languages that have certain basic structural properties. Similarly, even though it is obvious that not all people have the same values, the range of psychologically (and biologically) admissible values is limited. (This doesn't mean that there are a finite number of *specific* admissible values—it means that there are a finite number of *kinds* of admissible values.) It thus seems that (B) is false; for it doesn't seem correct to say that your innate psychological structure puts no limits on what kind of value system you can happily accept.

A stronger point seems to be warranted. If, as we've been arguing, a value system (or, possibly, a plurality of value systems) is internal to your emotional architecture, then there are limits not only to what kinds of value systems you can happily accept, but to what you kinds of value systems you can accept *at all.* Presumably, you cannot change (except within fairly narrow limits) how you emotionally react to situations. So supposing that, as we've been arguing, you can change your value system only by changing your emotional architecture, then you cannot (except in trivial respects)

change your value system at all. Of course, you can, in terms of your overt behavior, *comply* with a value system that doesn't make you happy. But if the view we've been defending is right, you don't *really* accept such a value system. If your emotions reflect your *actual* values, then you don't *really* accept a value system that brings you misery.

We may conclude that (B) is not true. It does not seem to be the case that one's innate psycho-biological endowment puts no limits on what kind of value system one can accept without jeopardizing one's chances of having a happy or otherwise good life.

Chapter 27

Law

It's widely held that laws are nothing other than mechanisms of coercion and, therefore, that law is primarily about power, and only secondarily, if at all, about morality. Nobody denies that laws and legal systems may *happen* to be good. But those who grant this often hold that it isn't *inherent* in the nature of law that legal systems must be good, their position being that any overlap between law and morality is incidental to what law is.

Is this right? Are legal systems primarily instruments of coercion, and only secondarily, if at all, ways of promoting the ends of morality? Or is it inherent in the nature of law that legal systems must embody a certain morality? The facts seem to demand that the first question be answered with a "yes" and the second with a "no." Fact: there are evil laws. Fact: laws are used to force people to do things they don't want to do and aren't ethically required to do. Fact: there are supremely immoral legal systems.

An analysis of law mustn't deny any of these obvious facts. Nonetheless, the concept of law *is* to be understood in terms of morality, and anything that fails to meet certain, non-trivial moral requirements *ipso facto* isn't a legal system.

If Smith and I were stranded on a desert island, and we thus both fell outside the scope of any legal system, I would have a moral, but not a legal, obligation to refrain from punching him. My legal obligation not to punch

Smith is therefore distinct from my moral obligation not to do so. But my legal obligation not to punch Smith is *a* moral obligation: it is a moral obligation to comply with a system of rules on which people depend for protections of their most basic rights—a system which builds highways and hospitals, prevents people from killing and stealing from one another, and holds physicians and other professionals accountable to certain ethical standards.

One may have a legal obligation to commit an immoral act. But, in such a case, one's legal obligation is not morally hollow. One's legal obligation to commit an immoral act is identical with one's moral obligation to comply with a system of the kind just described. Discharging that obligation *involves* committing an immoral act, but that obligation is not itself morally negative or morally hollow.

This last point demands elucidation. It is one thing to have an obligation to do X, and it is quite another to have an obligation the fulfillment of which *involves* doing X. A surgeon has an obligation to heal his patient. Fulfilling that obligation *involves* causing his patient to suffer. But a surgeon doesn't have an obligation to cause his patient to suffer. (In fact, he has an obligation to minimize his patient's suffering. A surgeon would be violating his professional obligations if he chose not to use anesthetic when there was no medical reason not to do so.) Similarly, fulfilling one's legal obligations may *involve* committing an immoral or morally neutral act. But it doesn't follow that legal obligations are ever immoral or amoral. If I have a legal obligation to do X, that means that I have a moral obligation the fulfillment of which *involves* my doing X. It does *not* mean that I have a moral obligation to do X. That is why a legal obligation to commit an immoral act may be morally wholesome. Hence the following argument is invalid—"given that one has legal obligations to commit immoral acts, it follows that legal obligations are not moral obligations"—and once this is seen, one of the more obvious obstacles to accepting a moralistic analysis of law has been removed.

Let us discuss a positive reason to accept such an analysis. A government is a success precisely to the extent that it serves the interests of those whom it governs. The same is therefore true of laws, given that laws are among the instruments of government. It follows that the concept of law is to be understood in moral terms, even though there are immoral laws.

Governments often behave in an unspeakably evil manner towards their own constituents; and often-times this misconduct involves, or even coincides with, the issuing of laws. On this basis, many have concluded that the concept of law is *not* to be understood in moral terms. But this line of thought involves a failure to make three important distinctions:

(i) The distinction between individual and institutional frames of evaluation;

(ii) The distinction between functional and purely factual frames of evaluation; and

(iii) The distinction between suspensions of protections of rights, on the one hand, and positive violations of rights, on the other.

The *personal* success of a head of government may not involve his serving the interests of his constituents, and it may even involve his thwarting those interests. But a head of government is a success *as* a head of state precisely to the extent that he serves the interests of his constituents. Given this, it follows that the objective of government—of *any* government—is to protect the interests of the governed. Stalin failed *as* a head of government precisely because he did not serve the interests of his constituents. It follows that the objective of the Soviet *government* was to serve the interests of its constituents—even though that may not have been the objective of Stalin or of any of the other senior officials composing that government. If a dictator issues a law that harms his constituents, he fails *as* a law-maker, even though issuing that law may help him achieve *personal* success of some kind—even though, for example, it may help him hold onto power. In general, one fails *as* law-maker if one issues a law that hurts those who are subject to it. This shows that the purpose of a legal system *per se* is to serve the interests of those who are subject to it—even though that may not be the objective of all, or even any, of the individuals who created that legal system. Distinction (i) thus removes one of the major reasons to reject a moralistic analysis of law.

Now let us discuss (ii). Even if everyone alive suddenly succumbed to liver disease, and each person's liver stopped removing toxins from his bloodstream, there would still be a significant sense in which livers were to be understood as removers of toxins. If one accepts a strictly fact-based frame of evaluation, one is blind to this obvious fact; and one is *not* blind to it if one accepts a frame of evaluation that allows for functional categories. An

analogue of this line of thought holds of law and government. Even if every government were thoroughly corrupt and exploitative, there would still be a significant sense in which governments were to be understood as protectors of rights, and in which the laws issued by governments were to be understood as attempts to provide protections of rights. Like hearts and livers, governments and the laws they issue are to be understood in functional, and not strictly statistical-empirical, terms.

The significance of distinction (iii) is to be understood in terms of the following argument. If a certain legal system doesn't prohibit anyone from violating Smith's rights, then Smith falls outside the jurisdiction of that system. There is legal obligation only where there is moral protection, and the moral core of one's legal obligations lies in this fact.

But there is an obvious objection to this argument. People have legal obligations under governments that harm them. Given this, it seems straightforwardly false to say that legal obligations under a government categorically presuppose receipt of moral protections from that same government. Here is where distinction (iii) becomes relevant. Suppose that Officer Smith beats me without cause. Even while I am being beaten, the government continues to protect my rights: anyone caught attempting to burgle my house or kidnap my children will be arrested; and it may even be other police officers who, acting in the name of duty, stop Smith from further beating me. Thus, Smith's misconduct is a case where a positive violation of my rights is superimposed on the continued existence of governmental protections of those same rights. The same thing is true of *any* case where one has legal obligations under a government that harms one. Thus, even though one may have legal obligations towards a government that harms one, it doesn't follow that legal obligation doesn't presuppose moral protection. So far as one thinks otherwise, one is failing to make distinction (iii).

Like the present author, Ronald Dworkin[426] argues that the concept of law is to be understood in moral terms. His argument is that judicial decisions are often made on moral, as opposed to narrowly statutory, grounds. This argument is less than probative. If the judge rules that the money is Smith's, then

the bank manager is acting illegally if he refuses to give Smith the key to the

safety-deposit box. If the judge rules that the money is *not* Smith's, then the bank manager is acting illegally if he *does* give Smith the key. The judge may or may not make his decision on the basis of a moral principle. In either case, the law is what the judge says it is. This is the antithesis of the idea that there is any *necessary* relationship between law and morality, even though it is obviously compatible with the idea that the two may sometimes overlap.

We've already seen why law is to be understood in terms of morality, notwithstanding that Dworkin's argument for this conclusion is less than cogent. But if our criticism of Dworkin is correct, it seems to follow the legal interpretation (so-called) is mere legislation, and thus isn't interpretation at all. But legal interpretation is *not* mere legislation, and this is perfectly compatible with the fact that the law is what the judge says it is.

A brief detour through the history of science may help to indicate the broader outlines of my argument for this. The layperson's concept of temperature has no application outside of what are, from a physicist's viewpoint, extremely narrow horizons. When physicists identified heat with molecular motion, they were not identifying the content of the layperson's concept of temperature. Rather, they were replacing the lay-person's concept with one that is consistent with it, within its narrow sphere of applicability, but is more precisely defined and also has a much wider sphere of application. Such principled extensions of existing concepts are referred to as "delineations" or "precisifications." Any judicial ruling is *ipso facto* an extension, not an identification, of existing law. But judicial rulings are *principled* extensions of the law—that is, they are delineations of it, and are not cases of law being created out of whole cloth. Just as some delineations of the concept of temperature ("a body's temperature is the sum of the velocities of its constituent molecules divided by the number of those particles") are truer to pre-theoretic data than others ("heat is dephlogistinated matter"), so some interpretations of law are truer to existing law than others.

Blackstone said that judges identify, and do not create, law. John Austin described Blackstone's position as a "childish fiction." Given the argument just outlined—and, in fact, independently of that argument—we must agree with Austin. But given that same argument, we must *dis*agree with the nihilistic view, held by the so-called "legal realists," that judicial

interpretation is mere legislation, and isn't answerable to objective standards.

Appendix 1

The Rudiments of Logic

1.1 The concept of an inference

To make an inference is to form a new belief on the basis of an old one. Here is an example. My friend Larry has been evicted from his apartment. I let him stay with me until he finds a new place to live. I notice that, shortly after Larry moves in, all of my money and valuables start disappearing. I also notice that purchases that I didn't make are appearing on my credit card bills. I know that nobody other than Larry had access to my money, valuables, and credit cards. So I infer that Larry has stolen my money and valuables and illicitly used my credit cards. (Thanks, Larry!)

Inferential knowledge is indirect knowledge. Non-inferential knowledge is direct knowledge. There is some reason to believe that *all* knowledge is inferential. (This issue is discussed in Chapter 10.) In any case, it's clear that some knowledge is *relatively* direct. My knowledge that I am now typing on a keyboard is *more* direct than my knowledge that heat is molecular motion.

1.2 Inductive inference vs. deductive inference

A ***premise*** is a belief that one either has or that one is willing to accept for argument's sake. An inference always begins with certain premises and ends with a conclusion that one accepts on the basis of those premises.

There are two kinds of inferences: deductive and inductive. An inference is deductive if the premises are intended to make the conclusion 100% likely to be correct. An inference is inductive if the premises are intended to ***support*** the conclusion. It's no easy matter to say what it is exactly for P to "support" Q. It's typically said that P "supports" Q if P, supposing it to be true, makes it probable but not certain that Q is true. This is what we'll say for the time being. (But, strictly speaking, it's false. The reasons for this are given in Section 1.9.)

Consider the following argument:

Premise 1: Smith drives a Rolls Royce.

Premise 2: Smith lives in a large house.

Conclusion: Smith is very wealthy.

Here the premises make it *probable* that Smith is rich, but they don't make it certain. It's possible that Smith stole the Rolls Royce or bought it on credit; and it's possible that, even though Smith lives in a nice house, he cannot afford it and is now in debt. So the premises don't make the conclusion certain, even though they obviously give it some weight; and the argument is therefore inductive.

Contrast that argument with this one:

Premise 1: Smith drives a Rolls Royce.

Premise 2: A Rolls Royce is a car.

Conclusion: Smith drives a car.

Here the premises, supposing them true, make it 100% likely that the conclusion is true. The conclusion can thus be deductively, as opposed to inductively, inferred from the premises.

1.3 The concept of entailment

If one statement can be deductively inferred from another, we say that the second *entails* the first. Let P be the statement that x is a square, and let Q be the statement that x has more than two sides. Since it is impossible for Q to be false if P is true, P entails Q.

1.4 The concept of confirmation

P confirms Q if the probability of Q, given P, is higher than the probability of Q, given not-P. Equivalently, P confirms Q if P, if true, fails to give Q a probability of 100% but nonetheless raises the probability of Q.

A *sufficient* (but not---see below---a necessary) condition for P's confirming Q is that Q be capable of being legitimately inductively inferred from P:

P: Smith drives a Rolls Royce, wears extremely expensive clothes, has an excellent credit rating, has an excellent employment history, all of his jobs being extremely lucrative ones, and he has no criminal record.

Q: Smith is wealthy.

P, if true, makes it sufficiently improbable Q is false that, given P, we may infer Q. That inference is not *deductive*, given that Smith might be an extraordinarily talented, but penniless con-artist.

Even though a *sufficient* condition for P's confirming Q is that Q be capable of being inductively P, it is not a *necessary* condition. For P to confirm Q it is *necessary* only that P increase Q's probability; it is not necessary that P increase Q's probability so much that, given P, Q is to be presumed true. Let P be the proposition: Smith is the most intellectually gifted person in Russia. And let Q be the proposition: Smith will find a cure for cancer. Obviously P *raises* the probability of Q. Equivalently, P raises the probability of Q, if only by a small amount, whereas not-P doesn't raise the probability of Q at all. But P doesn't Q enough weight for it to be the case that, given P, Q is to be presumed true. We'll clarify this point in Section 1.9.

True statements can *confirm* false ones, but they cannot *entail* them. Since "Smith drives a Rolls Royce and lives in a big house [etc.]" entails "Smith lives in a house," the second cannot possibly be false if the first is true. But since "Smith drives a Rolls Royce and lives in a big house [etc.]" merely confirms "Smith is wealthy," the second can be false if the first is true.

1.5 Validity vs. soundness

A *valid* argument is a deductive argument whose premises, if true, *in fact* give a probability of 100% to the conclusion. In other words, a valid argument is an argument whose premises are supposed to give 100% probability to the conclusion and whose premises *in fact* give 100% probability to the conclusion.

A valid argument can have false premises. For an argument to be valid it is necessary only that if the premises *were* true, then there *would be* a 100% chance that the conclusion is true. It may be that Smith doesn't drive a Rolls

Royce or any kind of car for that matter. But it's still the case that if Smith did drive a Rolls Royce, then, given that all Rolls Royces are cars, there would be no chance that Smith didn't drive a car.

A *sound* argument is a valid argument with true premises. (For example: JMK is a human; all humans are mammals; therefore, JMK is a mammal.)

A valid argument cannot possibly have true premises and a false conclusion, but it can have false premises and a false conclusion or false premises and a true conclusion. We'll discuss this in class.

1.6 Two kinds of induction

There are two kinds of inductive inference: ***induction by enumeration*** and ***inference to the best explanation.***

If, on the basis of the fact that you know of many x's that are y's and of no x's that aren't y's, you infer that *all* x's are y's or that the *next* x you encounter is a y, you've performed a case of induction by enumeration.

Here is an example. You've seen a million swans. They were all white. (Moreover, you *knew* in each case that what you were seeing was a white swan—you didn't think it was, for example, a beige duck.) You've never seen, or otherwise come to know of, any non-white swan. On this basis, you infer that *all* swans are white and, therefore, that, if you ever encounter another swan, it will be white.

If, in order to account for something of which you have knowledge, you posit the existence of something of which you don't have knowledge; you have made an *inference to the best explanation.*

All theories are cases of inference to the best explanation. All the theories put forth by Einstein, Darwin, Freud are instances of this mode of inference. Inference to the best explanation is a very powerful form of inference. (In fact, we'll see in Section 1.7. that it is the *only* form of non-deductive inference. So-called "induction by enumeration," to the extent that it isn't a spurious method of inference, collapses into inference to the best explanation.)

The conclusions of inferences to the best explanation are always *causal* statements. You posit the existence of X, which you don't know of, in order

to explain Y, which you do know of, because, in your judgment, X, if it existed, would *cause* Y to occur.

Example #3 is a case of inference to the best explanation. I infer that Larry stole my credit cards (etc.) because the relevant data is easy to explain on the assumption that he did, the reason being that his doing so would have *generated* that data, and hard to explain on the assumption that he didn't, the reason being that it's unclear what else could have generated that data.

To take another example: if you infer that a mouse has been eating the cheese in your cupboard, it's because a mouse's doing so would have generated the relevant data (e.g., the tiny little footprints inside the cupboard), and it's not clear what else could have generated that data.

In some cases, the conclusions of inferences to the best explanation aren't *themselves* causal statements. Our knowledge that water consists of H_2O molecules isn't itself a causal statement. Cause-effect relations hold between distinct entities (e.g., thunder and lightning). Water and H_2O aren't distinct.

That said, it is believed that water consists of H_2O molecules only because its being so composed would cause various things to happen that *do* in fact happen. So even though the conclusion of an inference to the best explanation need not *itself* be a causal statement, all inferences to the best explanation are causal inferences.

1.7 A fact about induction by enumeration (This section is hard. Skip on first reading.)

Induction by enumeration isn't nearly as powerful a form of inference as inference to the best explanation. Given only that every single one of the 5,000 first editions you've ever come across smelled like pipe tobacco, you cannot reasonably infer that all first editions smell like pipe tobacco. What if all of those first editions were the property of your pipe-smoking friend Larry? In that case, you'd have good to reason to suspect that their smelling like pipe tobacco had to do, not with their being first editions, but with their being things that had been around Larry. By the same token, so far as you are entitled to infer that all first editions smell like pipe tobacco, it's only to the

extent that you have reason to believe that a thing's smelling like pipe tobacco is *rooted* in its being a first edition. But that means that, if the data entitles you to infer that all first editions smell like pipe tobacco, it's only because it also entitles you to accept some inference to the best explanation (one to the effect that a thing's being a first edition is, for some reason or other, responsible for its smelling like pipe tobacco). In general, induction by enumeration is parasitic on inference to the best explanation. (In Chapter 12, it is shown how a failure to see this underlies the spurious belief that there is no non-deductive, inferential knowledge.)

1.8 Two kinds of entailment

Just as there are two kinds of inductive inference, so there are two kinds of deductive inference. In other words, there are two kinds of entailment: formal entailment and informal entailment.

An entailment is "informal" if the syntactic structures of the sentences involved are not what make the entailment go through. Consider the following entailment:

if Smith is tall and Jones snores, then Smith is tall.

This sentence has the form:

If P and Q, then P.

If a sentence has that form, it constitutes a valid inference. So # is made true by its form—that is, by its syntax.

But not all valid entailments hold in virtue of syntax. In fact, most do not. Consider the sentence:

^ If Smith weighs 1000 pounds, and Jones weighs 80 pounds, then Smith weighs more than Jones.

Obviously ^ is valid. But its form (i.e., its syntax) isn't what makes it true, since that sentence has the same syntax as:

^^ If Smith weighs 80 pounds, and Jones weighs 1000 pounds, then Smith weighs more than Jones,

which is obviously invalid.

2.0 Some notational conventions relevant to formal logic[427]

Statements of the form "if P, then Q" are known as "conditionals." The expression "→" stands for "if...then." So "P→Q" means "if P, then Q." Thus,

(1) Smith has four cars→Smith has more than one car

means that

(2) If Smith has four cars, then Smith has more than one car.

Thus, "P→Q" can be taken to mean that P implies Q. Thus "→" expresses implication.

As we will use the expression "→," the expression "P→Q" means that P *entails* Q, i.e., that it is *impossible for P to be true if Q is false.* In other words, "→" expresses what is sometimes referred to as ***strict entailment.*** (P strictly entails Q if, supposing that P is true, there is no way that Q can be false. In this work, "entailment" and "strict entailment" are synonymous expressions.)

This isn't how "→" is typically used. It is typically used to express a much weaker notion known as ***material implication.*** This term is defined below.

Biconditionals and "↔": Statements of the form "if P then Q, and if Q then P" are known as "biconditionals." When stated in non-artificial notation, "if P then Q, and if Q then P" is typically compressed into "P iff Q" (or "P just in case Q" or "P exactly if Q"). We'll compress it into: "P↔Q." So "Smith is a father↔Smith is a male parent" says the same thing as "Smith is a father iff Smith is a male parent."

Parentheses as indicators of scope: Parentheses are used to indicate how sentences are to be grouped together. Consider the sentence:

(3) "snow is white and grass is green→grass is green."

(3) is ambiguous between:

(4) it's the case that snow is white; and it's also the case that, if grass is green, then grass is green

and

(5) supposing that it's the case that snow is white *and* that grass is green, it follows that grass is green.

(4) and (5) have different meanings. (5) doesn't say that snow is white. Nor does it say that grass is green. It says that *if* it's the case that both grass is green and snow is white, *then* it's also the case that grass is green. (4), on the other hand, says (among other things) that snow is white. So (4) is false in universes where snow is green. But (5) is true in such universes. (In fact, (5) is true in every possible universe.) Since there are circumstances where (4) is false but (5) is true, they don't have the same meanings.

Modern symbolic logic gives us a way of avoiding the laborious *ad hoc* methods just used to disambiguate (3). To express (5), we say:

(3.5) (snow is white and grass is green)→grass is green.

And to express (4), we say:

(3.4) snow is white and (grass is green→grass is green).

In general, any two co-parenthetical sentences are more tightly bonded to each other than either is to anything that isn't co-parenthetical with those sentences.

Negation: ~P is the negation of P. Thus, "~(snow is not white)" is the negation of "snow is white." This means that "~(snow is not white)" says the same thing as "it is not the case that snow is white."

"~(snow is not white)" doesn't say that snow is black; nor does it say that snow is green. It says only that snow isn't white; everything else is left open.

The negation of a sentence isn't the same thing as the "opposite" of that sentence. In modern logic, the word "opposite" has no meaning.[428] The use of the "~" makes it easy to negate statements that would otherwise be hard to negate without prejudging important issues or using cumbersome constructions. Consider the sentence:

(%) If John likes to listen to Mozart, then Larry likes to listen to Brahms or the sky is blue.

In natural language one way to negate (%) is to put the expression "it is not

the case that'' in front of it. The result of this operation is:

($\%_1$) "it is not the case that if John likes to listen to Mozart, then Larry likes to listen to Brahms or the sky is blue."

But there's a problem. ($\%_1$) is ambiguous, as it could mean either:

($\%_2$) either the sky is blue or it is not the case Larry likes to listen to Brahms if John likes to listen to Mozart

or

($\%_3$) it isn't the case that the sky is blue; nor is it the case that Larry likes to listen to Brahms if John likes to listen to Mozart.

It takes work to come up with each of these two sentences. Each of the original two sentences had to be reparsed; and in each case this reparsing involved the use of methods that, being *ad hoc,* could be arrived at only through a creative insight and logical inference. Such Herculean measures shouldn't have been necessary, given that the objective was only to perform a purely grammatical operation.

This problem doesn't arise in our special symbolic notation. % can be unambiguously disambiguated in a mechanical, thought-free manner. One need only put a '~' before % and then enclose the part one wants negated in parentheses. Thus, the symbolic analogue of:

($\%_{2S}$) ~(Larry likes to listen to Brahms if John likes to listen to Mozart) or the sky is blue;

and the symbolic analogue of ($\%_3$) is:

($\%_{3S}$) ~(either the sky is blue or (Larry likes to listen to Brahms if John likes to listen to Mozart)).

In ($\%_{3S}$), the ***scope*** of the negation-sign is % in its entirety. (Were it not for the parentheses immediately flanking its second disjunct, ($\%_{3S}$) would be ambiguous between a conditional whose consequent was a disjunction and a disjunction one of whose disjuncts was a conditional.) In ($\%_{2S}$), the scope of the negation-sign is confined to the sentence "if John likes to listen to Mozart,

then Larry likes to listen to Brahms."

Thus, in our symbolic notation, parentheses are used is to make it clear how much scope is to be given to a given ***operator.*** An "operator" is an expression that, given one or more sentences, yields a new sentence. Examples of operators are: "it is not the case that," "or," "and," "because," "it is possible that," and "Fred believes that."

"Fred believes that" is an operator because, when given the sentence "snow is white," it yields a new sentence, namely: "Fred believes that snow is white." "And" is an operator because, when given the sentences "snow is white" and "grass is green," it yields the sentence "snow is white and grass is green." "Because" is an operator for the same reason *mutatis mutandis.* (Since "snow is white because grass is green" has a very different meaning from "grass is green because snow is white," "because" operates on *ordered* pairs of sentences.[429])

When an operator occurs in some sentence S, the scope of that operator is identical with the set of sentences out of which it is creating a new sentence.

Consider the sentence:

(i) "Jim is tired because he went jogging and his wife Sally is a very demanding person."

This is ambiguous. It could mean either:

(ii) Jim's wife Sally is a very demanding person; moreover, Jim is tired because he went jogging

or

(iii) There are two reasons why Jim is tired: first he went jogging; second, Jim's wife Sally is a very demanding person.

Thus, in English (or any other natural language), use of the expression "because" may yield a sentences that is ambiguous and, in addition, can be disambiguated only by successfully doing a certain amount of syntax-chopping and, therefore, logic-chopping.

But in our symbolic notation, one can disambiguate (i) without having to do any logic-chopping. One need only put parentheses around the two sentences that one wants joined by the "because." Thus, the symbolic

analogue of (ii) is:

(ii_S) (Jim is tired because he went jogging) and Sally is a very demanding person.

And the symbolic analogue of (iii) is:

(iii_S) Jim is tired because (he went jogging and Sally is a very demanding person).

Negation in relation to the use of parentheses: Parentheses make it clear what is being negated. (A) ~snow is white or grass is green

is ambiguous. It could mean either

(B) *either* it's the case that grass is green *or* it's the case that snow isn't white

or

(C) it isn't the case that (snow is white or grass is green).

(C) is equivalent with:

(D) snow isn't white and grass isn't green.

(B) isn't equivalent with (D), since (B), unlike (D), is true in a world where grass is green.

(A_1) ~(snow is white or grass is green)

says the same thing as (C), and

(A_2) ~(snow is white) or grass is green

says the same thing as (B). Thus, the parenthesis make it clear how much scope a given occurrence of "~" has—that is, they make it clear how much it is negating.

The modal operators—the box and the diamond: "□P" means "necessarily P" and "◊P" means "possibly P."

Note concerning the modal operators: The box and the diamond are to be given minimal scope. So "□P→Q" is *not* to mean that it's necessarily the case that P implies Q. Rather, it is to mean that, if it's necessarily true that P is the case, then it follows that Q is the case. "□(P→Q)," on the other hand, *does* say that it's necessarily the case that P implies Q.

The non-exclusive meaning of "or": We'll use "or" in the *non-exclusive* sense. So, as we'll be using it "either Smith is having dinner or Smith is in London" is consistent with Smith's being in London while having dinner. In order for "P or Q" to be true, all that is necessary is *one* of those two statements be true. So there are three circumstances under which "P or Q" is true: (i) P is true (but Q is false); (ii) Q is true (but P is false); (iii) P is true and Q is also true.

> In contemporary logic, both formal and informal, "or" is always used in the non-exclusive sense.

2.1 Some general principles relevant to formal logic

Deduction is truth-preserving: A true statement cannot entail a false one. In other words, a false statement cannot be validly inferred from a true one.

Deduction is transitive *(a corollary of the fact that it's truth-preserving):* R is a transitive relation if, supposing that x bears to R to y, and y bears R to z, x bears R to z. (The relation of being less tall than is transitive.) Deduction is transitive. A statement entails anything entailed by any one of its consequences. For example:

If Smith is an elephant, then Smith is a mammal.

If Smith is a mammal, then Smith has hair,

Therefore, if Smith is an elephant, then Smith has hair.

Induction isn't truth-preserving: A true statement may **confirm** a false one. The statement "Smith smokes eight packs of cigarettes a day" certainly confirms "Smith won't live to be a hundred." But there is still a chance that he will live to be a hundred.

Induction isn't transitive *(a corollary of the fact that it isn't truth-preserving):* A corollary of the non-truthpreserving character of induction is that induction is not transitive. In other words, given only that P confirms Q and that Q confirms R, it doesn't follow that P confirms R. Let P, Q, and R be defined as follows:

P: Smith is a morbidly obese man with a heart condition and terminal cancer who is due to be executed in five hours.

Q: Smith is a morbidly obese man with a heart condition and terminal cancer who is going to die soon.

R: Smith will soon die of some ailment related to his obesity, his heart-condition, or his cancer.

P confirms Q; Q confirms R. But P doesn't confirm R. P makes it a veritable certainty that Smith will be *executed* in five hours and, therefore, that he *won't* die for reasons having to do with his heart or weight or cancer-situation. So P disconfirms R, even though it confirms Q, which, unlike P, does confirm R.

Contraposition, conversion, inversion (not very important—skip on a first reading): These are operations that are performed on conditional statements (statements of the form "if P, then Q").

The ***contrapositive*** of (P→Q) is (~Q→~P).

Thus, the contrapositive of "if Bob is a dog, then Bob is a mammal" is "if Bob is not a mammal, then Bob is not a dog."

Statements are equivalent with their contrapositives. In other words, (P→Q)↔(~Q→~P). Bob's being a dog entails his being a mammal iff his being a non-mammal entails his being a non-dog.

The ***converse*** of (P→Q) is (Q→P). Thus, the converse of "if Bob is a human, then Bob is a mammal" is "if Bob is a mammal, then Bob is human."

As this example shows, statements are not generally equivalent with their converses. It's only when both ***antecedent*** and ***consequent*** are equivalent that a given conditional is equivalent with its converse. (In (P→Q), P is the antecedent and Q is the consequent.) Thus, the converse of "if Bob is a father then Bob is a male parent is true" is true, since antecedent and consequence are equivalent.

The ***inverse*** of (P→Q) is (not-P→not-Q). Thus, the inverse of "if Bob is a human, then Bob is a mammal" is "if Bob is not a human, then Bob is not a mammal."

As this example shows, statements are not generally equivalent with their inverses. It's only when both antecedent and consequent are equivalent that

conditional is equivalent with its inverse. This is because the converse of a proposition is equivalent with its inverse. In other words, (Q→P)↔(not-P→not-Q). The reason why they're equivalent is that the second is the contrapositive of the first. (As we know, contraposing a conditional produces an equivalent conditional.)

A sentence-level operation is ***truth-preserving*** if, when given a true sentence as input, it cannot possibly yield a false sentence as output. Contraposition is truth-preserving. Conversion and inversion are not.

Exercise: Prove that, given any conditional sentence S, the contrapositive of S is equivalent with S. Further, prove that it is not the case that, given any conditional sentence S, the converse of S is equivalent with S. Finally, prove that, given any conditional sentence S, the inverse of S is equivalent with the converse of S.

Material vs. strict vs. formal entailment:

Logicians distinguish between "formal" and "material" implication, and they distinguish each from "strict" implication (which is usually referred to as "strict entailment").

"Strict implication" is synonymous with "entailment." As previously stated, we'll use "→" to express strict entailment. So, as we'll mean it, "P→Q" says that P entails Q, i.e., that Q cannot possibly be false if P is true.

P formally entails Q if the sentence ‹if P, then Q› is an instance of an ***open-sentence*** that is true for all values of its free variables. Thus, a formal truth is a truth that is an instance of an open-sentence that is true for all values of its free variables. "2 = 2" is a 'formal' truth, because it is an instance of the open-sentence

"x = x," which is true for all values of its variables.

Formal implication is a *kind* of strict implication. Any case of formal implication is a case of strict implication, but not *vice versa.* The reasons for this are given in Chapters 1, 6, and 18 and also in Section 4.5 of the present.

"P materially implies Q" is true if any one of the following three conditions is met:

(i) P is false

and Q is false;

(ii) P is false
and Q is true;

(iii) P is true
and Q is true.

It is only if

(iv) P is true and Q is false

that P does *not* materially imply Q.

If P is any false statement P, and Q is any statement at all, "P is true *and* Q is false" is false. For this reason, any false statement materially implies all statements.

Material implication isn't a form of implication at all. So, while it's a fact that, for any statement Q, any given falsity "materially implies" Q, that fact is an innocuous consequence of a linguistic convention that, although it concerns the word "implication," has nothing to do with implication at all.

Material and strict entailment may hold between propositions or between expressions. But formal entailment holds only among expressions. It is typically held to hold among sentence-types. My strongly held

view is that it holds only among sentence-tokens. (See Chapter 4 for a defense of this claim.) This is the view of Peter Strawson, Jon Barwise, and John Perry.[430] But, for the sake of brevity, we'll use the term 'sentence' to stand for sentence-type and sentence-tokens.

Modality: "Modal" means "having to do with necessity or possibility." Possibility and necessity are properties of statements, not objects. Any statement of the form □***P*** ('it's necessarily the case that P') or ◊***P*** ('it's possibly the case that P') is a modal statement.

"Necessary" and "possible" are interdefinable. "Necessary" can be defined as "not possibly not," and "possibly" can be defined as "not necessarily not." Thus, □P iff ~◊~Q.[431]

Statements that affirm *bearing relations,* whether deductive or inductive, are modal. Thus, causal statements are modal statements, since they have the

form "□(if P, then Q)."

Causal necessity isn't logical necessity, of course. Given that (i) I threw gasoline on the fire, it is *causally* necessary that (ii) the fire flared up. But (ii) isn't a logical consequence of (i). In other words, it can't be known through meaning-analysis or syntax-chopping that (ii) cannot be false if (i) is true.

2.3 Some rudiments of formal logic[432]

(Note: during a first reading, skip the commentary.)

1. For any x and any y, ((x = y)→(for any property F, (x has F↔y has F))).

This principle is known as "Leibniz's Law."

Some paraphrases: Given anything x and anything y, if x is the very same thing as y, then x has a given property just in case y has that property.

If x is the very same thing as y, then anything true of the one is true of the other.

A *necessary* condition for identity is sameness of properties.

Commentary on Principle 1: In Chapter 16, it is argued that, although this principle is clearly true for "synchronic" identity-clams, it is clearly false for "diachronic" identity claims.

2. For any x and any y, (given any property F, ((x has F↔y has F)→(x = y)))

This principle is known as "the identity of indiscernibles."

Some paraphrases: Given anything x and anything y, if it the case that x has a given property just in case y has that property, then it follows that x is the very same thing as x.

If x has no properties that y lacks and y has no properties that x lacks, then x is the very same thing as y. A *sufficient* condition for identity is sameness of properties.

Commentary on Principle 2: Though it has been questioned, this principle is clearly correct. Given any object x, x has the property of being identical with x. So, given any object y, if y has every property that x has, then y has the property of being identical with x and is therefore itself identical with x.

But, while it's true that anything having that property is identical with x, it's also trivial. And my suspicion is that those who claim to be asking

whether the identity of indiscernibles is correct are really asking a distinct but related question, namely: given an object y that has every property had by x, setting aside those properties that are *trivially* had by anything identical with x (e.g., the property of being identical with x), is y identical with x?

That question, unlike the question of whether the identity of indiscernibles is correct, isn't trivial. That said, the answer to it does seem, quite clearly, to be "yes." If, at any given moment, x is in the same place as y, has the same mass, is subjected to the same forces, and so on, then surely x is identical with y. (But I'm not

100% convinced by my reasoning here. Exercise: If you think that the identity of indiscernibles is correct, prove it. If you think that it's false, prove that.)

Commentary on Principles 1 and 2: The identity of indiscernibles is often said to be the "converse" of Leibniz's Law.

This isn't correct. For P to be the converse of Q, it is necessary that P be a conditional statement and that Q also be a conditional statement. Leibniz's Law isn't expressed by a conditional statement, and neither is the identity of indiscernibles.

Each is expressed by a ***universal generalization.*** In each case, that universal generalization is to the effect that each member of an infinitely large class of conditional statements is correct. But neither is itself expressed by a conditional statement, and neither is itself a conditional proposition.

Leibniz's Law says that:

(LL) for any x and any y, if x is identical with y, it's impossible for x to lack any property had by y or vice versa.

LL is not itself a conditional statement. For the same reason *mutatis mutandis,* the identity of indiscernibles isn't given by a conditional statement. Therefore, neither principle is the converse of the other.

3. P or ~P. (The law of excluded middle.)

Any given statement is either true or false.

Commentary: Is principle 3 correct? That depends on what the word "statement" is taken to mean. If it's taken to mean the same thing as "sentence," then 3 is false. The sentence (in other words, the sentence-type) "I am tired" is neither true nor false. (Certain *utterances* of it are true and certain utterances of it are false.)

If the word "statement" is taken to mean the same thing as "sentence-utterance" (or, more generally, "***sentence-token***"), then, once again, 3 is false. If, while pointing to an empty space, I say "that elephant is in terrible pain,"[433] what I've said is neither true nor false. There is no such elephant.

Therefore, there is no elephant x such that I am saying that x in terrible pain. If I am saying anything at all, there is some elephant x such that what I am saying is true iff x is in terrible pain. Since there is no such elephant, I'm not saying anything; my utterance is neither true nor false.

If the word "statement" is taken to mean the same thing as "proposition," then 3 *is* correct—or so I believe. (Many disagree. See the entry titled "ambiguity vs. vagueness" in Appendix 2.) What we're seeing here is that the so-called truths of symbolic logic—the "laws of logic"—aren't true at all. They're statement-*forms,* not statements, and therefore aren't true or false. And it isn't easy to find interpretations of those statement-forms that validate them. In other words, it isn't easy to figure out what sorts of constants must replace the variables occurring in these "laws" if true sentences are to result. This suggests that modern symbolic logic (a.k.a. mathematical logic, a.k.a. formal logic) isn't logic at all. It's actually a small and rather unimportant branch of a discipline known as "model theory." In Section 3.4, we'll see why, given this fact, it follows that formal logic has little to do with the way in which any actual or even possible being reasons.

4. ~(P and ~P). (The law of non-contradiction.) No statement is both true and false.

5. (P and (P→Q))→Q. (Modus Ponens.) Supposing that P, and supposing that P entails Q, it follows that Q.

Here is an example, which we'll refer to as "SD" (short for "Smith is a dog"):

Premise: Smith is a dog.

Premise: If Smith is a dog, then Smith is a mammal. Conclusion: Smith is a mammal.

Commentary: Consider the following argument, which we'll call "SD#":

(1) Smith is a dog.

(2) All dogs are mammals. (In other words, if a given thing is a dog, then that thing is a mammal.) Therefore, (3) Smith is a mammal.

SD# is *not* an instance of modus ponens. To be an instance of modus ponens, an argument must have the form:

(1*) P.

(2*) If P then
Q. therefore,

(3*) Q.

Let's refer to *this* argument form as "AF*." If, in AF, we replace "P" with "Smith is a dog" and we replace "Q" with "Smith is a mammal," the result is SD, not SD#. There are no sentences such that, if the sentence-level variables in AF are replaced with those sentences, the result is SD#. Therefore, SD# is not an instance of modus ponens.

SD# is a valid argument. But it isn't an instance of modus ponens. It's an instance of the following argument form:

(1#) A has property P.

(2#) For any x, if x has P, then x
has Q. Therefore, (3#) A has Q.

Let's refer to this argument form as "AF#." SD# makes a statement *about* SD. SD# says that SD is an instance of modus ponens. If S is a statement that says of some argument A that A is an *instance* of modus ponens, S is not *itself* an instance of modus ponens. Thus, SD# is not itself an instance of modus ponens.

6. (P→Q) and ~Q) →~P. (Modus tollens.)

Premise: If Smith is a dog, then Smith is a mammal.

Premise: Smith isn't a mammal.

Conclusion: Smith isn't a dog.

7. (P and ~Q) →~(P→Q)

Deduction is truth-preserving. No true statement entails a false one.

Explanation: If P is *actually* true and Q is *actually* false, there is obviously a *possible* circumstance where P is true and Q is false. (After all, what is actual must be possible. The impossible isn't actual.) An immediate consequence is that P doesn't entail Q.

8. ((P→Q) and (Q→R))→(P→R) (If P entails Q and Q entails R, then P entails R.)

Equivalent with: (P→Q)→((Q→R))→(P→R)). (If P entails Q, then P entails R provided that Q entails R.)

Entailment is transitive. A statement entails anything that its consequences entail. This is a corollary of Principle (7).

9. P→~~P. (The *law of double-negation.*)

Premise: Smith is a professor.

Conclusion: It's not the case that Smith is not a professor.

10. P→(P or Q).

No false disjunction contains a true disjunct.

11. ~(P or Q)↔(~P and ~Q).

A disjunction is false iff all of its disjuncts are false.

Exercise: Are principles (10) and (11) equivalent? Defend your answer.

Note: there is a trivial sense in which any two (correct) logical principles are equivalent. A logical principle is, by definition, one that holds in all possible worlds. Since any two logical principles hold in all possible worlds, they are all equivalent. (That is, they are all ***model-theoretically*** equivalent.) But in this context, interpret the term "equivalent" more narrowly. Two sentences are "equivalent," as we'll be using this word in this particular context, if they say the same thing, the only differences between them being grammatical. (A good test of whether two statements are equivalent in this narrower sense is whether, given some natural language L (e.g., French, Arabic, Finnish), there is some sentence S of L that is clearly the translation of the one *but not the other.*

12. ~P→~(P and Q).

No true conjunction contains a false conjunct.

13. ~(P and Q)↔(~P or ~Q).

A conjunction is false iff at least one of its conjuncts is false.

Exercise: Are principles (12) and (13) equivalent. Defend your answer. (Remember what we said a moment about what the word "equivalent" is to mean in this section.)

14. ((P or Q) and ~P)→Q

Either Jones is a plant or Smith is in the house. Jones isn't a plant.

Therefore Smith is in the house.

Exercise: Are (13) and (14) equivalent? Defend your answer.

15. ~((~P and Q)→~(P→Q))

A false statement may entail a true one.

Explanation: "Paris is in France and Spain is north of Sweden" entails "Paris in France." Any given true statement is entailed by each of the infinitely many false conjunctions of which it is a conjunct. For that reason, and others, a false statement may entail a true one. (But a true one cannot entail a false one.)

Exercise: Are (14) and (15) equivalent (in the previously discussed sense of "equivalent")? What about (13) and (15)? Defend your answer.

16. ~((P→Q)→(Q→P))

Propositions aren't always consequences of their own consequences. Conversion isn't truth-preserving. "Paris is in France and Spain is north of Sweden" entails "Paris in France," but not *vice versa.*

17. P→P

Any given proposition entails itself.

Explanation: P entails P iff it isn't possible for (P and ~P) to be true. It isn't possible for (P and ~P) to be true. (The law of non-contradiction forbids it.) Therefore, P entails P.

18. ~Q→((P→Q)→~P)

Anything that entails a falsity is itself false.

19. (Q→(P and ~P))→~Q.

Anything that entails a statement *and* its negation is false.

Explanation: No true statement entails a false one. (Principle 18.) (P and ~P) is false. (Principle 4.) So for any proposition P, no true statement entails (P and ~P).

20.(P→~P)→~P

A statement is false if it entails its own negation.

Explanation: Any given statement entails itself. (Principle 17.) So if a statement P entails its own negation, it follows that P entails (P and ~P). (P and ~P) is false. (Principle 4.) Therefore, a statement is false if it entails its own negation.

21. (P and ~P)→Q

There is nothing that isn't entailed by a contradiction.

Explanation: There is no possible circumstance where (P and ~P) is true. Therefore, for any proposition Q, there is no possible circumstance where Q is false *and* where (P and ~P) is true.

Some principles of modal logic:

1M. P→◊P. Whatever is actual is possible.

2M. □P→P. Whatever is necessary is actual.

3M. □P→◊P. Whatever is necessary is possible.

Given that squares must have four sides, they do have four sides.

Commentary: This shows that there is a difference between the possible and the contingently true. The contingently true doesn't have to be true; but the possible is sometimes necessarily, and therefore non-contingently, true.

4M. (P→Q)↔~◊(P and ~Q).

P entails Q iff P can't be true if Q is false.

Explanation: This is a definition, not a principle. To say that P entails Q *is* to say that it's impossible for P to be true if Q is false.

5M. □(P and Q)↔(□P and □Q).

A conjunction is necessarily true iff each conjunct is necessarily true. Explanation: If P isn't necessarily true, there is some possible circumstance where it's false and where, therefore, (P and Q) is false, no matter what Q is. If (P and Q) is necessarily true, there is no circumstance where (P and Q) is false or, therefore, where either P is false or Q is false.

6M. □~(P or Q)↔(□~P and □~Q).

A disjunction is necessarily false iff each disjunct is necessarily false.

Explanation: Each of P and Q must be false in every possible world in order for ~(P or Q) to be false in every possible world; and there's no way that ~(P or Q) can be true in any possible world if each of P and Q is false in every possible world.

7M.(□P and ◊~Q)→~(P→Q).

Nothing that can be false is a consequence of anything that cannot be false.

Explanation: If P is necessarily true (true in all circumstances), whereas Q is possibly false (false in some circumstances), there are circumstances where P is true and Q is false (i.e., P doesn't entail Q).

8M. ~((◊P and ◊Q)→◊(P and Q)).

Conjoining two possible statements may yield a necessarily false statement.

Explanation: "On June 12, 2009, JMK did not at any time leave the state of Virginia" is possible, and so is "On June 12, 2009, JMK did at some point leave the state of Virginia." But they're not ***compossible*** (i.e., they cannot both be true).

9M.(◊P and □~Q) → ~(P→Q)

Nothing that can be true entails anything that cannot be true.

Explanation: If P is true in some possible circumstance, whereas Q is true in none, there is a possible circumstance where P is true and Q is false, which means that P doesn't entail Q.

10M. □P→(Q→P)[434]

Given any necessarily true statement, there is no statement that doesn't entail it.

Explanation: If P is necessarily true, there are no possible circumstances where P false. *A fortiori,* there is no possible circumstances where, for any proposition Q, P is false *and* Q is true.

11M.~□P→(P→Q)

Given any statement at all, be it true or false, there is no necessarily false statement that doesn't entail it.

Explanation: If P is necessarily false, there is no possible circumstance where P is true. *A fortiori,* there is no possible circumstance where, for any proposition Q, P is true *and* Q is false.

12M. □(P→Q)→(□P→□Q).

A statement is necessary if it's entailed by a necessary conditional with a necessary antecedent.

> Exercise: 13M can be seen as a generalization of Modus ponens. Explain why this is so. Exercise: Is "(□P→□Q)→ □(P→Q)" a correct principle? Explain why or why not.

13M. □P→□□P. ("S5.")

Whatever is necessary is necessarily necessary.

Explanation: Point #1: If ~□□P, then there is some possible circumstance where ~□P. By contraposition, if there is *no* possible circumstance where ~□P, then □□P is true. Point #2: If ~□P, then there is some possible circumstance where ~P. By contraposition, if there is *no* possible circumstance where ~P, then □P is true. Point #3: Given the first two points, it follows that, if □P, then □□P. Conclusion: S5 is correct.[435]

14M: ◊P→◊◊P.

What is possible is possibly possible.

3.0 Some meta-logical principles

3.1 Model theoretic entailment

P model-theoretically entails Q just in case there is no coherently conceivable scenario where P is true and Q is false.

Thus, *1 + 1 = 2* model-theoretically entails *triangles have three sides,* even though the two propositions do not, apart from that, have anything to do with each other.

This last point is important. There isn't anything about shapes, let alone triangles specifically, in *1 + 1 = 2;* and there isn't anything about addition in *triangles have three sides.* So even though, technically, each entails the other,

there is no significant sense in which the one can be *inferred* from the other —no sense in which a knowledge of the one yields a knowledge of the other.

Of course, somebody who knows that 1 + 1 = 2 will probably also know that triangles have three sides, but he won't know the one on the basis of the other. By contrast, somebody who knows that x is a case of true belief could potentially know that on the basis of the fact that x is knowledge.

Commentary Model-theoretic entailment is an over-valued notion that doesn't correspond at all well to the conception of entailment embodied in our inferential practices. Nonetheless, because, in the early '60's, some branches of philosophy were (wrongly) thought to be successfully assimilated to a branch of mathematics known as "model theory"—the purpose of this branch being to study model-theoretic entailment—philosophers now tend to treat model-theoretic entailment as *the* fundamental kind of entailment—when, for the reasons just stated, it is in fact an entirely artificial notion.

Model-theoretic entailment isn't really a kind of entailment. It is a *theory* of entailment. Like other scientists, logicians model data. Where logic is concerned, the data in question are our intuitions concerning bearing-relations—they are to the effect that certain things entail (the negations of) certain other things. We know intuitively that

(K_1) "x is knowledge"

entails

(K_2) "x is a true belief."

Logicians want to make it clear what it is that we know in knowing this; they want to make it clear what it is for the one statement to entail the second.

Here is the proposal that they came up with:

(MT[436]) for P to entail Q is for there to be no possible circumstance in which the first is true and the second is false.

Why did logicians choose MT when trying to model our intuitions concerning entailment? Because MT is *extensionally* correct. In other words, for any statements P and Q, P entails Q iff there is no possible circumstance where P is true and Q is false.

But even though there are no counter-examples to it, MT is false. It's obviously in virtue of facts about it's

structures of K_1 and K_2 that the former entails the latter. The fact that there are no models in which K_1 is true and K_2 is false is a mere *consequence* of that structural fact.

There is a related point. MT defines entailment in terms of "possible." This makes it viciously circular. P and Q are compossible (in other words, (P and Q) is possible) iff neither entails the negation of the other. We understand the concept of possibility *only* to the extent that we can understand it in terms of entailment (and other bearing-relations). So unless we're willing to take the radically implausible view that "possible" is a ***primitive expression,*** MT defines "entailment" in terms of itself. (A primitive expression is one that doesn't consist of other expressions. Examples of such expressions are "red" or "sweet.")

To make MT non-circular, we must rid it of the word "possible." Thus modified, MT is:

(MT*) for P to "entail" Q is for there to be no circumstance in which the first is true and the second is false.

In MT*, 'circumstance' must be mean *actual* circumstance. If it means anything else, MT* will be circular, like MT.

MT* is simply wrong. There is no circumstance in which the moon is made of cheese. And there is no circumstance in which I have ten cars. So there is no circumstance in which "the moon is made of cheese" is true and in which "JMK has ten cars" is false. So, according to MT*, the former entails the latter. But the former doesn't entail the latter. (The former "materially implies" the latter. But that's irrelevant, since material implication isn't entailment.)

So model-theoretic entailment isn't entailment at all. It's what a wrong *theory* of entailment says that it is.

Some of the principles stated in Section 2.3, not all of them, go through only if "entails" (i.e. "→") is taken to refer to model-theoretic entailment.

3.2 Formal entailment

In some cases, one statement's following from another can be determined entirely on the basis of the structures of those sentences. For any statements P and Q, ‹P and Q› entails ‹P.› For any statement P, ‹not not P› entails ‹P.›"not-not-P entails P" is correct.

If it can be determined that Q follows from P entirely on the basis of the structural properties of those sentences, P ***formally entails*** Q.

For awhile, it was thought that all entailments could be formalized—that is, that whenever one statement entails another, it is entirely because of the syntactic or structural properties of those sentences. That viewpoint is simply false. Why it is false is explained in Chapter 1, Section 6.0.

3.3 Formal entailment a relation between expressions

Only expressions have syntax. So, since P formally entails Q only if P and Q have certain *syntactic* properties, the relation of formal entailment holds between *sentences* (or sentence-tokens), whereas the relations of model-theoretic entailment holds between propositions, the same being true of "informal, non-model-theoretic entailment," which I will now define.

3.4 The one, genuinely important kind of entailment: the informal, non-model-theoretic kind

Let P1 be the statement *John knows that 1 + 1 = 2,* and let Q_1 be the statement *John believes that*
1 + 1 = 2. P_1 model-theoretically entails Q_1. In other words, there is no coherently conceivable scenario where P_1 is true and Q_1 is false. But the entailment relation that holds between P_1 and Q_1 is not *just* an instance of model-theoretic entailment; it is also an instance of another, more significant kind of entailment. This is made clear by the fact that there is a marked difference between the relation that holds between, on the one hand, P_1 and Q_1 and, on the other hand, the entailment-relation that holds between:

P_2: 1 + 1 = 2

and

Q_2: triangles have three sides.

The difference is clear. Given the information contained in P_1, it can be deduced that Q_1 is true. Q_1 is, in some important sense, "implicit in" or "contained in" P_1. But Q_2 isn't implicit in P_2. There isn't anything concerning triangles in P_2—not implicitly or explicitly.

When we make inferences, they are *not* of the strictly model-theoretic kind. Rather, they are of the kind where the statement that is inferred is, in some way or other, contained in the statement from which the inference is made. So the inference from:

P_1: John knows that 1 + 1 = 2

to

Q_1: John believes that 1 + 1 = 2

is an instance of the conception of entailment that is embodied in our actual inferential practices. But that inference isn't formal; it isn't like the inference from *snow is white and grass is green* to *snow is white:* the formal properties of P_1 and Q_1 leave it open whether Q_1 follows from P_1. (After all, those sentences are formally identical with

P_3: John is sure that 1 + 1 = 2

and

Q_3: John doubts that 1 + 1 = 2.

P_1 is syntactically just like P_3, and Q_1 is syntactically just like Q_3. But there is no legitimate inference from P_3 to Q_3; and since those sentences are syntactically just like P_1 and Q_1, respectively, it follows there is no legitimate, purely syntactic inference from P_1 to Q_1.

Although, as previously discussed, P_1 model-theoretically entails Q_1, P_1 also entails Q_1 in some other, much more robust sense. In Kuczynski (2007), I put forth a theory as to the nature of this other, more important sort of entailment.

3.5 The ampliative character of deduction

It is often said that, if P entails Q, there is nothing in Q that is not also in P. So deduction (i.e., the identifying of entailments) is, according to this, "non-ampliative"—the consequent doesn't "amplify" the antecedent (i.e., it doesn't say anything not said by the antecedent).

This position is obviously false where model-theoretic entailment is concerned. (*Triangles are shapes* model-theoretically entails *Socrates is identical with Socrates,* but there is nothing in the content of the first that has anything to do with Socrates.) And it's obviously false where formal entailment is concerned. ("Smith is tall" formally entails "either Smith is tall or grass is green," but there is nothing about grass in "Smith is tall.")

3.6 The most important kind of entailment: the non-formal, non-model theoretic kind

The
statement:

(P) 1 + 1 = 2

mode-
theoretically
entails

(Q) the interior angles of a Euclidean triangle add up to 180°.

This is all very well. But it's patently obvious that you can't *infer* the truth of P from that of Q, or *vice versa.* In general, model-theoretical entailment is inferentially useless. A *precondition* for knowing that P entails Q is that you know, first, that P cannot be false and, second that Q cannot be false. But if you know this, there's no need to infer the one from the other. What this shows is that knowledge of model-theoretic entailments are

parasitic on some other, much more fundamental sort of entailment. And what *this* shows is that knowledge of model-theoretic entailments have no significant role in the thought-processes of any possible cogitator.

Unlike model theoretic entailment, formal entailment isn't inferentially useless. But our ability to make formal inferences is parasitic on our ability to make informal inferences. Actually, so-called "formal" inferences *are* informal ones.

Formal entailment is a relation that holds between expressions (more specifically, sentences or sentence-tokens). This means that we have to know what sentences mean to make inferences that are licensed by formal entailments. To infer

(S_1) "snow is white"

from

(S_2) "snow is white and grass is green,"

I have to know what those two sentences mean, at least up to a point. I have to know that "and" doesn't mean what is in fact meant by "or." I also have to know that "snow is white" and "grass is green" are complete sentences—that they aren't drivel, like "bunga berka der."

So in order to infer S_1 from S_2, I have to know the truth of some proposition of the form:

(S_3) *given* that S_2 means such and such and that S_1 means thus and such, it follows that S_1 is true, given that S_2 is true.

But S_3 is not formally true. First of all, it's a proposition, not a sentence. But even the corresponding sentence (the result of putting the verbiage in question in quotation marks) isn't formally true. It's an analytic, logical, *informal* truth.

Second reason: Formal truths, and thus formally true entailments, are *instances* of universal generalizations that, although analytically true, are not themselves formal truths. The sentence:

(S_4) “if snow is white and grass is green, then snow is white”

is formally true. But why? Because

(S_5) for all sentences[437] P and Q, ‹(P and (P→Q))→P›,

though true, isn’t *formally* true. It *instances* (e.g., “if snow is white and grass is green, then snow is white”), but not it itself. This is because it has the same form as:

(S_6) For no sentences[438] P and Q, ‹(P
and (P→Q))→P›,

which is false.

Some logicians[439] say that, because S_5 has an “all” in the place where S_6 has a “no,” they *don’t* have the same form. But that’s just a way of saying that they don’t have the same form because, if they did, then S_5 wouldn’t be a formal truth anymore. It’s a fact that, in any sense of the expression “same form” that isn’t circularly constructed to validate preexisting theories, S_5 and S_6 have the same form. They certainly have the same syntactic form, at least as linguists use the word “syntax.” The rules of sentence construction involved in the one case precisely parallel those used in the other. It’s a simple fact, then, than S_5 and S_6 do not, in any significant sense, differ in form and that S_5, though true, isn’t formally so.

Given that formal truth is parasitic on informal truth, the reason being that instances of formal truths are instances of informal ones, entailment cannot ultimately, or even proximately, be understood in formal terms and, in addition, only a fraction of the inferences that we make are formal. (This is setting aside the fact noted a moment ago that so-called formal reasoning is really informal reasoning in disguise.)

3.7 Entailment not a limiting case of confirmation

In some contexts, entailment can be treated as a limiting case of confirmation: P’s entailing Q can be seen as P’s confirming Q to the highest

possible degree.[440]

It's tempting to see entailment as maximal confirmation. But that isn't what it is. For P to entail Q (in other words, for it to be analytic that Q follows from P) isn't for P to confirm Q to the highest possible degree. Entailment is a logical notion. Confirmation is an explanatory notion. For P to confirm Q to be a high degree is for (*P and not-Q*) to be counter-explanatory—in other words, it is for (*P and not-Q*) to be false *unless* deeply rooted presumptions about how the world works are uprooted and replaced. For P to entail Q is for (*P and not-Q*) to be counter-conceptual, i.e., it is for the structures of concepts, unassisted by presumptions as to how the world works, to prohibit the falsity of Q, given the truth of P. So P's confirming Q to the highest possible degree is different from P's entailing Q.

3.8 Entailment both an intersentential and an interpropositional relation

Sometimes the word "entailment" refers to a relationship that holds between propositions (sentence-*meanings*), and sometimes it refers to a relation that holds between expressions of some kind. Some logicians hold that those expressions are sentence-*types*. Others, myself included, hold that those relations may hold between sentence-types *and also* between sentence-*tokens*. (Peter Strawson (1950) was the first to distinguish token-entailment from type-entailment. In so doing, he performed a major service for philosophy.)

With rare exceptions (Strawson being one of them), authors who discuss "entailment" don't make it clear whether they are talking about a relationship that holds between propositions, sentences, or neither.

In this book, unless there is an indication to the contrary, "entailment" will refer to a relationship that can hold either propositions *or* between expressions.

3.9 The irrelevance of formal logic to thought

For reasons that were discussed in Chapters 1, 7, and 18, any assistance that knowledge of formal logic can provide one in the way of acquiring

knowledge is parasitic on one's knowledge of informal analytic truths. This has nothing to do with the idiosyncrasies of human psychology. It's an epistemological consequence of a strictly logical point. Given any open-sentence (or open-proposition) all of whose instances are true, for example:

(1) ‹(P and (P→Q))→Q›,

the *fact* that it's true for all its instances is a consequence of the fact that some informal truth holds—for example:

(2) given any two propositions P and Q, ‹(P and (P→Q))→Q›).

It must be stressed that (2) is an informal truth. (2) has the same form as the false statement that:

(2#) given no two propositions P and Q, ‹(P and (P→Q))→Q›).

Nothing that has the same form as a false statement is formally true. Therefore, (2), though true, isn't formally so. (2#) is a statement *about* a class of formal truths; but, like most statements *about* formal truths, it is not itself a formal truth.

Bearing this in mind, consider the following statement, which, unlike (2), *is* formally true:

(3) "if snow is white, and (if snow is white then snow is not pink), then snow is not pink)."

There are two different reasons why one might accept (3). One is that one knows it to be an instance of (1) and one accepts it for that reason. But in that case, one is accepting (3) on the grounds that it's a consequence of (2). Since (2) is informal, one's acceptance of (3), under those circumstances embodies knowledge of an informal truth.

The same thing holds if one's acceptance of (3) is based on one's understanding of the concepts meant by the expressions composing it. Suppose that your reason for accepting (3) is that, given your understanding of the concept expressed by "if...then...," you know that it would be absurd to deny (3). In that case, your acceptance of (3) is based on your knowledge that, *the structure of that concept being what it is,* any proposition of that form must be true. Thus, your acceptance of (3) is based on your acceptance of (2). Given that (2) is an informal truth, your acceptance of (3) under those

circumstances constitutes informal knowledge. There is no circumstance, therefore, under which your knowledge of (3) is any sense formal. The concept of "formal" (or "mechanical") thought is therefore an incoherent one.

4.0 Models

An open-sentence is an expression that contains a free variable, and is therefore neither true nor false, but is otherwise just like a sentence. So ‹x is an even number› is an open-sentence, since it contains a variable ("x") where there should be a "2" or a "4" or some such. It often happens that, given a set of one or more open-sentences, we wish to find some way of replacing the variables with constants that yield true sentences. For example, let S_1 be the set containing the following two open-sentences (I'll henceforth omit the use of quasi-quotation marks):

(1) x is an even number that is greater than zero.

(2) x_2 is less than 20.

If we assign the number two or the number four to x, the result is that both sentences are true. So each of these assignments *validates* S_1.

In general, given a set of open-sentences, an assignment of constants to the free variables validates that set iff two conditions are met. First, that assignment turns *each* open-sentence in that set into an actual sentence. Second, every sentence is true.

An assignment of constants to a set of open-sentences is an *interpretation* of that set. An interpretation of such a set *validates* it iff, under that interpretation, every sentence in that set comes out true.

Let S_2 be the set containing only the following open sentences:

(1) x is an odd number.

(2) x^2 is greater than

(3) √x is irrational.

If we assign the number nine to x, (1) and (2) come out true; but (3) does not. But if we assign the number 13 to x, (1)–(3) all come out true. (It is assumed that the universe of discourse is the set of numbers. This assumption will continue to be made until further notice.)

An assignment of constants to such a set of open-sentences is an ***interpretation*** of that set. Not all such sets are validated by all interpretations of them, as we just saw.

False interpretations must be distinguished from partial interpretations. A partial interpretation is one that is incomplete but is correct as far as it goes. Let S_3 be the set containing only the following open-sentences:

(1) x is a phi.

(2) x is larger than y.

(3) y is a phi.

(4) y is a psi.

If we assign the numbers 10 and 8 is to x and y respectively, and the property of being even to "phi," we've partially interpreted S_3, since, under that interpretation, (1)–(3) come out true. But since we haven't yet assigned a meaning to "psi," we haven't completely interpreted it. We *have* completely interpreted that set of open sentences if we make the further stipulation that for a thing to be a psi is for it to have the property of being less than 9. In fact, supposing that we make that additional stipulation, we have provided a complete interpretation of them that validates them. If we had instead stipulated that for a thing to be a psi is for it to be less than 5, we would have provided a complete interpretation of those open sentences that failed to

validate them.

Oftentimes, the interpretations that are of interest are very abstract ones. Let S_4 be the set containing just the following open-sentences:

(1) x is a phi.

(2) x bears R to y.

(3) y is a psi.

(3) R is "reflexive" (any given thing bears R to itself).

(4) y bears R to z.

(5) z is a phi.

(6) R is "transitive" (if a bears it to b, and b bears it to c, then a bears it to c).

Here's an interpretation of S_4. x, y, and z are, respectively, the numbers two, three, and four. Phi and psi, respectively, are the properties of being even and odd. R is the less relation of being less than or equal to. y is the number.

Question: Does the following set (S_5) have an interpretation that validates it? If so, what is it?

(1) x is a phi.

(2) x bears R to y.

(3) Anything that bears R to anything bears R to itself.

(4) y bears R to z.

(5) Nothing bears R to anything that has phi.

Some sets of open-sentences are true under *all* interpretations of them. An example is the set containing: "x is identical with x" and "if, for some phi, x has phi and y does not have phi, then x is not identical with y." Others (e.g., any set containing "x is not identical with x") are true under none.

Any ***interpretation*** of a set of the first kind is a formal truth. Thus "JMK is identical with JMK, and anything identical with a professor is identical to nothing that isn't a professor" is a formal truth.

Appendix 2

Important Terms and Principles

Algorithm: A fixed procedure for carrying out a task The rules that we learn in grade school to multiply, add, etc., multi-digit numbers are algorithms. By *formalizing* inferences, logicians create algorithms for determining whether, given two statements, one of them follows from the other.

A problem with algorithms is that what one has to know *in order to know* whether a given algorithm is applicable to a given task is often much heftier than what one would have to know to carry out that task in an *ad hoc* manner. In this fact lies the doom of attempts to algorithmize—or, to use the more popular term—"mechanize" thought.

Ambiguity vs. indexicality: An ambiguous expression is one that is assigned meaning by two or more different rules. So "dumb" is ambiguous, since there are two rules that assign meaning to utterances of it. One of those rules is: given an utterance of "x is dumb," that utterance is true just in case x cannot speak. The other is: given an utterance of "x is dumb," that utterance is true just in case x is not intelligent.

An "indexical" is a context-sensitive expression. For an expression to be context-sensitive is for there to be some *one* semantic rule that assigns different meanings (or referents) to it, depending on the context. An example of such expression would be the pronoun "I." That expression isn't ambiguous. Some one semantic rule assigns meaning to any two tokens of it. (That rule is: given any token t of "I," t refers to the person who produced t.) But, that rule being what it is, different tokens of "I" refer to different people.

Ambiguity vs. vagueness: Whereas an expression is ambiguous if it's assigned meaning by more than one semantic rule, an expression is "vague" if, supposing it unambiguous, it is assigned meaning by *no* well-defined semantic rule, and if, supposing it ambiguous, one of its disambiguations is assigned meaning by no well-defined semantic rule. Someone with zero hairs is definitely "bald," and someone with a million hairs (provided that they're suitably located and have the requisite thickness) definitely is "not bald." But there are many people with an intermediate number of hairs with respect to

whom neither "bald" nor "not bald" is clearly applicable. Since, therefore, the semantics of "bald" is given by a rule that is undefined for these intermediate cases, "bald" is vague.

Like many other expressions, the word "dumb" is both vague and ambiguous, since it is assigned meaning by two distinct semantic rules, neither one of which is entirely determinate.

Commentary: There is a raging debate as to what vagueness is. Some hold that it is a property *only* of words, thoughts, and the like. Those who hold this are said to have an *epistemic* view of vagueness, and are sometimes known as "epistemicists." ("Epistemic" means "having to do with knowledge.") This is because they see vagueness as a property, not of the thing known, but of our knowledge of it.

Others hold that vagueness is a property of the thing known, and not, at least not merely, of our knowledge of it. Those who hold this are said to believe in "objective" vagueness, the reason being that they believe vagueness to be a property of objects, as opposed to our methods of representing them.[441] According to this view, there may be no fact as to whether Pat is a male or not or as to whether Pat is kind or not, etc. If this is right, then the law of excluded middle (any given proposition P is either 100% true or 100% false) is false.

Here is my view. There is no objective vagueness. "What's out there is out there," as my former colleague Chris Buford once put it. Talk of vagueness in the world is projective. We're projecting deficiencies in our representations of the world onto the world itself. Vagueness is a property of beliefs, symbols, and other representations.

Here's why I say to this. To say of a predicate P that it's vague is to say that there is, or at least could be, some object x such that ‹Px› isn't true and ‹not-Px› isn't true. There's nothing mystifying about this sort of vagueness. Symbols have the meanings we give them. If we decide that Px is true if x is 1, 2, or 3, and false if x is 4, 5, or 6 then ‹P7› isn't true and neither is ‹not-P7.› But this isn't because there is some *proposition* that is neither true nor false. It's because, given how P has been defined, ‹P7› doesn't *encode* a proposition and neither does ‹not-P7.› For a representation or symbol to be "vague" is simply for it to be under-defined.

But what would it be for a symbol-*meaning* to be vague? But what would it be, for example, for a sentence-meaning (a proposition) to be vague?

Propositions are individuated by their entailment-relations. In other words, a proposition's identity is a function of what it entails and also of what entails it. If proposition P_1 entails P_2, but P_3 does not, or P_2 entails P_1 but not P_3, then P_1 and P_3 are *ipso facto* different propositions. Thus, the very idea of a proposition whose entailment-relations are indeterminate is a non-starter. This means that propositions always have determinate truth-conditions. After all, for proposition to have such and such truth-conditions *is* for such and such propositions to entail it. Thus, a proposition's truth-conditions are either fulfilled or they aren't. There is no other possibility. Thus, there are no indeterminate propositions, and the law of excluded middle (*every proposition is either 100% true or 100% false*) is correct. See "entailment," "propositions."

Antecedent and consequent: In *if P, then Q*, P is the antecedent and Q is the consequent. So in *if snow is cold, then snow is not hot*, the statement *snow is cold* is the antecedent, and *snow is not hot is the consequent.* See "conditional."

Analytic truth: "Squares have four sides" is analytic, since it makes no sense to suppose that squares might fail to have four sides. "Anything that is literate is animate" is analytic since it makes no sense to suppose that anything inanimate should be literate. In general, a statement is analytically true if its negation is incoherent.

See "coherent," "empirical truth," "entailment," and "meaning vs. entailment-relations."

Atomic proposition: A proposition that isn't molecular (e.g., *John snores*). An atomic sentence is a sentence that expresses such a proposition (e.g., "John snores"). See "molecular proposition."

A priori: Knowledge is a priori if one has it, not in virtue of any observations that one has made, but in virtue of one's innate cognitive structure of one's mind. See "empirical knowledge."

A posteriori: Knowledge is a posteriori if it isn't a priori. See "*A priori.*"

Axiom: A statement that, in some context, is assumed to be true and therefore isn't argued for.

Axiom (second definition): A statement-form that, in some context, is assumed to have a true interpretation. See "interpretation."

The axiom of comprehension: This is the principle that, given any property, there is a set containing all and only those objects that have that property. There is a set containing all and only those things that are people (i.e., that have the property of being a person). There's a set containing all and only those things that are acorns (i.e., that have the property of being an acorn). And so on.

The axiom of comprehension seems self-evident. But Bertrand Russell discovered that it has self-contradictory consequences and is therefore false. The set of people is not a member of itself, since no person is a set. But some sets do seem to be members of themselves. The set of sets would seem to be such a set.

In light of these points, consider the property of being a set that contains all and only those sets that are not members of themselves. Let S be the set that contains all and only those things that have property. Does S itself have that property? In other words, is S itself a self-member? If it is, then it isn't, since it contains only sets that aren't self-members. But it if isn't, then it is, since it contains every set that is a self-member. Thus:

(i) S has the property of being a self-member

entails its own negation, and so does

(ii) S doesn't have that property.

Given the axiom of comprehension, one of those two statements must be true. Since the axiom of comprehension therefore entails a contradiction, it is false. How exactly it must be modified is a matter of debate. See "axiom," "property," and "set."

Axiom of extensionality: Sets are identical if and only if they have exactly the same members.

Axiomatic system: The set of statements consisting of certain axioms along

with the statements that follow from them. See “axiom.”

Axiomatization: To axiomatize a discipline is to identify a small number of propositions and a small number of inference-rules such that all of the results of that discipline follow from those propositions by means of those inference rules. An axiomatization of a discipline is what results when it is successfully axiomatized. See “inference rule.”

Backtracking counterfactual: *See “counterfactual.*”

Bound variable: See “open-sentences.”

Causal law: A law to the effect that, if one thing occurs (e.g., fire), some other thing must occur (e.g., smoke).

Causal inference: An inference based on knowledge of a causal law (e.g., I see smoke; I know that there is smoke only if there is fire; so I infer that there is, or was, a fire).

Causal series: A series of events such that, given any two of them, one is the (in)direct cause of the other; in other words, a single series of events any one of which is (in)directly caused by some one state of affairs.

Like all series, causal series don’t branch. This is a tautology. A ‘series’ is defined to be a non-branching structure.

Of course, just a river can fork, so a causal series can branch out. In fact, the forking of a river is a special cause of the branching out of a causal series. But for a river to fork is for it to become *two* rivers (supposing that the two tributaries don’t later merge). It isn’t for *a* river to become two independent streams of water. Similarly, for a causal series to branch is for it to become multiple series. It is not for *a* series to comprise independent tributaries.

Class: Synonymous (in this work, though not in others) with “set.” See “set.”

Coherence and incoherence: An “incoherent” statement is one that undermines itself. Consider the statement:

(BC) “Bill has four cars, but he doesn’t have more than one car.”

BC is self-defeating. Its own meaning prevents it from being true. Therefore it is “incoherent.”

It's clear that *any* statement to the effect that someone had four cars while having no more than one car would be incoherent. So the concept of such a person is one that couldn't be actualized; and any statement to the effect that it was actualized would be incoherent. Such concepts are described as "incoherent." (In this context, the term "concept" is synonymous with the term "condition," and the expression "to actualize a concept" is synonymous with the expression "to fulfill a condition.")

Concepts are conditions. To actualize a concept is to fulfill a condition.

"***Ceteris paribus***": Synonymous with "other things being equal" and "holding all other factors constant." *Ceteris paribus*, a person with money is more likely to be happy than a person with no money. In other words, given two *otherwise comparable* people, one of whom has money, the other of whom does not, the former is more likely to be happy than the latter. But if the former has a painful chronic illness and his dreams have been all been shattered (etc.) whereas the former is in the pink of health and he feels himself to be on the cusp of success, the latter will probably be happier than the former.

Coherently conceivable circumstance (or scenario): A circumstance is coherently conceivable just in case one is not guilty of self-contradiction in virtue of believing it to hold. So even though Kerry is not the U.S. President in 2007, the proposition *Kerry is U.S. President in 2007* is not self-undermining—it isn't like *x is knowledge but not belief* or *1 + 1 = 3.* So while all beliefs requiring the existence of circumstances that are *not* coherently conceivable are false, not all false beliefs require the existence of such circumstances. See "coherence and incoherence."

Compatibilism: The doctrine that there can be personal freedom in a deterministic world. "Incompatibilism" is the doctrine that compatibilism is false. See "determinism" and "incompatibilism."

Compatible: P and Q are compatible statements if they can both be true. Thus, P and Q are compatible if, supposing that the one is true, it doesn't follow that the other is false. "So *JM is a philosophy professor*" is compatible with "*JM is more grouchy than Steve Carey*," since they can both be true. See "entailment" and "incompatible."

Complex expressions vs. simple expressions: Some expressions are simple (e.g., "Smith"), while others are complex (e.g., "the man that the dog bit"). All sentences are complex—that is, they consist of multiple expressions.

At first, it appears as though there are exceptions to this. Consider the sentence "leave!" This seems simple. But it isn't; that is because there is a hidden (or "phonetically unrealized," meaning that there is no overt, acoustical, or visual representation of them in that word) occurrence of "you"—that sentence means: *you*—leave. Also, the verb leave has various

semantic properties (e.g., tense, mood) that, although they are not phonetically marked, are nonetheless present, and make it a complex expression.

Genuinely simple expressions are known as “morphemes.”

Compositionality: According to the *principle of compositionality*, the meaning of a complex expression is a function of the meanings of its constituents; and the referent of a complex expression is a function of the referents of its constituents.

The meaning of:

(TY) “the author of *War and Peace* was friendly to the author of *Crime and Punishment*”

is different from that of

(DO) “the author of *War and Peace* was mean to the author of *Crime and Punishment.*”

Why do TY and DO have different meanings? Because “friendly” and “mean” don’t mean the same thing. Replacing “friendly” with “mean,” or *vice versa*, turns a sentence having one meaning into a sentence having a very different meaning. This shows that what TY means *depends* on what “friendly” means. Semanticists express this by saying that TY’s meaning is “a function of ” what “friendly” means. What DO means is a clearly function of what “mean” means. In general, what expressions mean is a function of what the expressions composing them mean. See “special compositionality,” “function,” and “functor.”

Compound sentence: Same as a molecular sentence. Thus, a compound sentence is one that consists of other sentences (e.g., "snow is white and grass is green"). See "molecular sentence."

Conceptual role semantics (CRS): the view that meaning is use, i.e., for two expressions to mean the same thing is for them to be used in the same way and, therefore, that for a given expression to have this, as opposed to that, meaning is for it to be used in this, as opposed to that, way.

In his (1963) paper "Some reflections on language games," Wilfrid Sellars clearly puts forth a version of CRS. But it's quite definitely present in Wittgenstein's book *The Philosophical Investigations*, which was completed in the late 40's. In his (1994) book *Making it explicit*, Robert Brandom advocates CRS. Nowhere in that 900 page book, or in any writings of advocates of CRS, is the existence of any of the following reasonable positions so much as acknowledged, let alone addressed:

(i) A knowledge of a word's meaning is at least part of what guides a person to use it in the way she does. When the waiter asks you what you would like to eat, you say "I'd like a bowl of clam chowder," and not "I'd like a bowl of rotten entrails." Why do you say "clam chowder" instead of "rotten entrails"? Because you know that, given what it is that you want, "rotten entrails" has the wrong meaning and "clam chowder" has the right meaning. Of course, how an expression is used is likely to *change* its meaning. If enough people use the word "restive" to mean *tired*, it may well come to have that meaning. But until that time comes, those people are *mis*using that expression; and their usage of it, being discrepant from its actual meaning, can't possibly be constitutive of it.

(ii) Mere noises aren't linguistic expressions. It is only noises coupled with meanings that are such expressions. In a world where there are no semantic conventions, a noise just like the noise you make when you say "snow is white" would mean nothing. If, in such a world, the gases leaving a volcano happened to make that noise, they would mean nothing. (In fact, they'd mean nothing in *our* world, except at most in some derivative sense.) Since, therefore, noises must be coupled with meanings in order to qualify as linguistic expressions, it makes no sense to say that what an expression

means is a function of how it is used. After all, there is no expression to be used *until* the noises (and ink-marks and so forth) have been coupled with meanings. Therefore, CRS is false.

Of course, there are different views as to what meanings are, and there are different views as to what it is for noises to be coupled with them. But relative to any such view, CRS is false.

The contention that is closest to CRS that isn't obviously false is the trivial contention that what noises (and ink-marks and the like) mean is a (partial) causal consequence of what people do with them. (This suggests that CRS is based on a failure to distinguish the concept of a logical consequence from that of a causal consequence.)

(iii) Expressions that haven't ever been used have determinate meanings. There are infinitely many sentences of English that haven't been used but that have determinate meanings. If this weren't the case, then nobody could ever utter a new sentence of English, at least not without making some *ad hoc* extension to the English language. But previously unuttered sentences are uttered all the time; and no sooner are they uttered than they have determinate meanings.

According to CRS, what an expression means is determined by how it is used. So if CRS is right, an expression that hasn't been used *ipso facto* has no meaning.

To parry this, advocates of CRS would have to say the following. "Where some expressions are concerned, meaning is fixed by use. But where other expressions are concerned, use *doesn't* fix meaning, at least not single-handedly. Some *other* factor is at work."

In making this concession, the advocate of CRS is admitting that CRS is false. Setting this aside, let us ask: what is this other factor?

The answer: syntax. The meaning of "Bob kicked the can" is a function of (i) what each of its components means (in other words, it's a function of what is meant by each of "Bob," "kick," and so on); and it's also a function of (ii) the syntactic rules of English (in other words, it's a function of the rules that assign meanings to complex expressions on the basis of the meanings of their constituents).

Thus, the meaning of a complex expression is *not* determined by its use,

and is determined by existing semantic and syntactic conventions. So CRS is false, if taken as an analysis of complex expressions. Earlier, we saw why it's false if taken as an analysis of simple expressions.

(iv) If the meaning of a sentence S were determined by its "conceptual role"—that is, by what people inferred from S, when they thought it correct —any inference that people made from S would *ipso facto* be constitutive of S's meaning. If enough people thought that "Stalin the dictator of the Soviet Union from 1923–1953" entailed "somebody who was over 6 ft tall was the dictator of the Soviet Union from 1923–1953," the former would entail the latter. But even if the latter were correct (which it isn't since Stalin was barely over 5 ft tall), it wouldn't *follow* from the first sentence.

In their (2001) book *The Compositionality Papers*, Jerry Fodor and Ernie Lepore put forth several blistering criticisms of CRS. In that work, any gaps in the criticisms of CRS just put forth are very much filled in.

Conditional: A conditional statement is one that has the form *if P, then Q.* Conditional statements play a vital role in logic, since logic is the discipline that attempts to determine what follows from what, and any statement to the effect that one statement follows from another is expressed by a conditional.

Confirm: For P to confirm Q is for P to provide support for Q. (P: 'Smith's fingerprints are found in the murder-weapon used to kill Jones.' Q: "Smith killed Jones." P supports Q.) Confirmation is defeasible, meaning that, even if P confirms Q, some other statement might provide more support for not-Q than P provides for Q. (R: "There is a video-tape of Brown killing Smith, then pressing the hand of Smith, who is at the crime-scene but unconscious, around the murder-weapon, and then leaving it at the scene.") ***Entailment*** is not defeasible. A consequence is that P cannot, if true, entail Q without Q's being true. See "confirmation."

Confirmation: P confirms Q if, supposing that P is true, it's possible that Q is false but it's more likely than not that Q is true. "Smith is a Harvard professor" confirms "Smith is not a complete moron," but it doesn't entail it. See "confirm" and "entailment."

Conjunct: In *P and Q*, each of P and Q is a "conjunct." So *snow is white* is

one of the conjuncts of *snow is white and grass is green.* See "conjunction."

Conjunction: A conjunction is any sentence composed of two sentences that are joined by an "and" or a "but" or an "although"—or any other connective whose purpose is to indicate that *both* of the sentences thereby joined are true.

Connective: A "connective" expression is one that, given one or more sentences, enables a new sentence to be formed. So, "if " is a connective since, given the sentences "grass is green" and "snow is white," it yields the sentence "if grass is green, then snow is white." All conjunctions ("and," "or," "but") are connectives. So is "not." This is because "snow is not white" says the same thing as "it is not the case that snow is white," which is what results when the expression "it is not the case" is coupled with "snow is white."

"It is possible" (or "possibly") is a connective, since it yields "it is possible that John is home" when given the sentence "John is home."

"Sally believes" is a connective since it yields "Sally believes that John is home" when given the sentence "John is home."

"And," "if," and "or" are *two*-place connectives, meaning that they form sentences out of pairs of sentences. "Not," "it is possible that," and "John believes that" are *one-place* connectives, since they form sentences out of single sentences.

Commentary: A connective can be thought of as a sentence-level adjective. That is, it can be thought of as an expression C such that, given a sentence S, C yields a new sentence C(S) (read: "C of S") such that C(S) attributes some characteristic to the proposition (or open proposition) expressed by S. For example, "Fred knows that" is a connective. And if you say "Fred knows that snow is white," you are attributing a property to the proposition that snow is white; you are saying of that proposition that it has the property that Fred knows it to be true. "Because" is a connective, and if you say "snow is white because grass is green," you are seeing of the propositions that snow is white and grass is green that the former is a causal consequence of the latter. One last example: "for all x" is a connective, and if you say "for all x, x is a mammal if x is a whale," you are attributing a characteristic (that of being

correct) to each of *infinitely many* propositions. (That is why, even though "for all x" and other quantifiers don't operate on sentences, they nonetheless qualify as connectives.

Connectives attribute properties to propositions. This is what quantifiers do. In fact, quantifiers do it on an even larger scale than other connectives. Consider the quantifier "for all x." When coupled with the open sentence "x is tall," this yields a sentence ("for all x, x is tall") that says of an infinitely large class of propositions (viz. those having the form x is *tall*) that each of its members is true. (By contrast, "Fred believes that," when coupled with "Jim is tall," yields a sentence ("Fred believes that Jim is tall") that attributes a property (that of being believed by Fred) to only *one* proposition.) That is why, even though quantifiers operate on open sentences, whereas other connectives operate on *actual* sentences, quantifiers are generally treated as connectives.

A sentence can have more than one connective. An example of such a sentence would be:

(*) "either Jim is in Idaho and Larry is in Idaho with Jim, or Larry is in Delaware and Fred is with Larry in Delaware."

The main connective of (*) is "or." Why is it the *main* connective? (*) attributes a property to the set containing the propositions *either Jim is in Idaho and Larry is in Idaho with Jim* and *Larry is in Delaware and Fred is with Larry in Delaware.* The property that it attributes to that set is the property of containing at least one true member. The other connectives occurring in (*) *don't* attribute properties to that pair of propositions; they attribute properties to propositions composing the propositions composing (*). That is why none of those connectives is the main connective. I leave it to the reader as an exercise to produce an explicit definition of the term "main connective" on the basis of these remarks.

Consequentialism: The doctrine that it is entirely in virtue of what its effects are that an act is morally right or wrong. See "deontology" and "utilitarianism."

Consistent/inconsistent: If two statements are compatible, they are "consistent" with each other; otherwise they are inconsistent with each other.

"Consistent" and "compatible" mean the same thing. See "compatible."

Consistent/inconsistent (in connection with axiomatic systems): When an axiomatic system is described as "consistent" what is meant is that, given any two statements composing it, they are consistent with each other. An axiomatic system is "inconsistent" if it doesn't satisfy this condition. See "axiom" and "axiomatic system."

Content: The content of a statement or a belief is what must hold for it to be correct.

Content-externalism: Let X and Y be two creatures that are qualitatively identical *except* for the fact that the causes of X's current condition aren't qualitatively identical with those of Y's current condition. According to content-externalism, the content of X's thoughts and perceptions may, in virtue of that fact, be different from the contents of Y's corresponding thoughts and perceptions. See "semantic externalism."

Contingent: A proposition is contingent if there are coherently conceivable circumstances where it is a true and also conceivable circumstance where it is false (e.g., *Mary owns a BMW*).

Counterfactual: a statement of the form: *if P were the case, then Q would also be the case* (e.g., *if Kennedy hadn't been assassinated, he would have been reelected in 1964*).

A backtracking counterfactual is one of the form: "if P were the case at time t, then Q would have been the case at time t*, "where t* is *earlier* than t." For example, "if Jim now had alcohol on his breath, then it would have been the case that he had been drinking earlier today."

Counterfactual analysis of causality: The position that x caused y to occur just in case, if x hadn't happened, y wouldn't have happened. According to this view, for the button's being pushed to cause the elevator to come is for it to be the case that the elevator wouldn't have come if the button hadn't been pushed. In Chapter 17, it is shown why this view is false.

Contextual definition: See "denotative vs. ostensive vs. descriptive vs. contextual definition."

Defeasible: A tendency or principle is "defeasible" if it can be overridden.

Definite description: Expressions of the form ‹the phi›, where phi is in the singular, are known as "definite descriptions." So "the inventor of Velcro," is a definite description, as is "the inventor of bifocals." See "quantifier."

Define/definition: To define an expression is to say what it means. Expressions are defined, not objects. "Bill Clinton" is defined, not Bill Clinton.

Definition by abstraction: In some cases, it can be said what it is for two things, x and y, to share some property P without using any expression that refers to P. In such cases, it is possible to translate statements in which expressions referring to that property occur into statements in which such expressions don't occur; and when this is done, the property in question is thereby defined in abstraction.

For example, two lines have the same direction (i.e., share the property of pointing in a given direction) if they don't intersect. Thus, the direction D of a given line L may be defined as the class of all lines L* that don't intersect with L, and "y has direction D" may be seen as saying the same thing as "y doesn't intersect with L."

Demonstrative: Some indexicals must be accompanied by an act of demonstration—an act of pointing, or some equivalent—if they are to succeed in picking out their target. If, while in the presence of 10 people, I say to you "that person is a bore," what I've said is ambiguous. But if, under the same circumstances, I utter the same sentence while pointing to Charlie, then my utterance of "that person" picks out Charlie, and my utterance is not ambiguous. An utterance of "that person" refers to the person who is salient in the context of utterance. There is thus a *rule* that assigns referents to tokens of "that person," that rule being *if "that person" is uttered in a context in which x is a uniquely salient person, then that utterance picks out x.* What that refers to is therefore systematically context-sensitive, that being why it's an indexical. And sometimes an act of pointing is needed to make somebody be uniquely contextually salient, that being why it's a demonstrative, as opposed to a mere indexical. Other examples of demonstratives are "this pain," "these little creatures," and "that aardvark." See "indexical."

Denotative vs. ostensive vs. descriptive vs. contextual definition: For an expression to be defined *ostensively* is for it to be defined by means of an act of pointing. If, having been asked who "Ludwig" refers to, I say "*that guy*," while pointing to Ludwig, I've defined "Ludwig" ostensively.

Ostensive definitions are always cases of *denotative* definitions, but not *vice versa.* For an expression to be defined denotatively is for it to be defined by identifying some entity that it picks out. This is how proper names are defined. "Socrates" is defined denotatively—you learn what "Socrates" means by being told who it refers to. But since he's no longer around, it isn't defined ostensively.

Not all expressions can be defined denotatively, and such expressions must be defined *contextually.* "Something" doesn't pick anything out. Therefore, it can't be defined denotatively. Its meaning is given by the rule that, for any phi, ‹something has phi› is true exactly if phi is instantiated. Other examples: The meaning of "every person" is given by the rule that, for any phi, ‹every person has phi› means that the property of being a human non-phi isn't instantiated. The meaning of "not" is given by the rule that, for any sentence S, not-S is true just in case S is false. The meaning of "no aardvark" is given by the rule that, for any phi, ‹no aardvark has phi› is true exactly if the property of not being both an aardvark and a phi is universally instantiated. See "define/definition."

Commentary—A principle concerning contextual definition: Whenever an expression is defined contextually, as opposed to denotatively, whole sentences containing it must be reparsed if their meanings are to be made clear. We just saw this with "something." Its meaning is given by the rule that, for any property psi:

(SP) ‹something has psi›

means

(PS) the property of being a psi is instantiated.

PS doesn't have the same form as SP. PS is a wholesale reparsing of SP; and the reason is obviously that "someone" is defined contextually. The same is true of all expressions that are defined contextually.

For any psi, ‹something has psi› attributes a property to another property,

and not to an individual. (In analytic philosophy, the term "individual" refers, not only to people, but to *anything* that isn't itself a property—to rocks, trees, galaxies, etc.) "Someone is tall" attributes the property of being instated to the property of tallness.

Additional commentary: All denotative definitions are contextual definitions, but not *vice versa.* To say that, for some object x, "Smith" refers to x is to say that, for any phi, ‹Smith has phi› is to the effect that x has phi. So in defining "Smith" by identifying its referent, one *is* giving its meaning by saying what is meant by whole sentences of the form ‹...Smith...› This principle is defended in Chapter 7.

Denote: As we'll use them, the terms "denote" and "refer to" are synonymous, the same therefore being true of "denotation" and "referent" are synonymous. See "refer."

Deontology: The doctrine that it is in virtue, not of what its effects are, but of whether it embodies due

regard for one's ethical duties and for other people's rights that an act is morally right or wrong. See "consequentialism."

Determinism: The doctrine that *nothing* is uncaused. Everything that happens/has happened/will happen *has to happen* and, moreover, *has to happen in the exact way in which it happens.*

Alternate definition of determinism: how the world is at any point fixes how the world will be, in every respect, at all later times. See "indeterminism."

Descriptive proposition: A proposition is descriptive if it isn't normative. In other words, a proposition is descriptive if it, supposing it is correct, says how things are, not how they should be. *Killing is frequent, especially during war* is not a normative statement, since it merely says how things are, not how they should be. See "proposition" and "normative proposition."

Disjunct: In *P or Q*, each of P and Q is a "disjunct." See "disjunction."

Disjunction: A disjunction is any sentence composed of two sentences that are joined by an "or." For example, *either it's raining or somebody turned the sprinkler on.*

Duhem-Quine thesis: The thesis, which was advocated by W.V.O. Quine (1908–2001)[442], that no theory does either a better job or a worse job than any other theory of modeling any given body of data. In other words, given any body of observations B and given any two theories T and T*, T does no better and no worse a job than T* of modeling B.

Quine's argument was that, given any datum D that appears to confirm T and to disconfirm T*, one can make D disconfirm T and disconfirm T* by tinkering with one's background assumptions.

Quine's thesis is either empty or it's false. Scientific theories aren't analytically entailed by the observations that support them. Induction isn't deduction. This is an obvious triviality. So, assuming that what Quine is saying isn't trivial, it must be to the effect that no theory does a better job or a worse job than any other theory of accounting for any given body of data. To account for a body of data is to show that the occurrence of each datum composing it is non-anomalous. Thus, Quine's thesis, supposing it non-trivial, is to the effect that no theory does a better job than any other theory of removing causal anomalies from any given body of data.

Thus interpreted, Quine's thesis is clearly false. The acne of each of the millions of acne-sufferers who takes medication X gets ten times worse within ten minutes of the time at which said acne-sufferer takes X. Theory T entails that anyone who has acne who takes X will immediately be cured of his acne forever. Rival theory T* entails that theory T is false. According to Quine, by tinkering around with background assumptions, T models the observational data as well as T*. But that's clearly false.

Quine says that, in situations like this, we should "plead hallucination." In other words, we should assume that the observations in question, most of which are had by people who do not otherwise have any history of hallucinations, are hallucinations. If we make this assumption, we're stuck with a huge number of otherwise non-existent anomalies. We're stuck having to explain, or to regard it as incapable of explanation, why it is that people who never had hallucinations before all of a sudden started hallucinating; and we are also stuck having to explain/regard as inexplicable why it is that people who did have hallucinations before all of a sudden started having acne-related hallucinations that, in terms of their content and in terms of their

etiology, are completely unlike the hallucinations they had previously.

Empirical knowledge, analytic knowledge, synthetic, a priori, a posteriori: Empirical knowledge is knowledge that is based at least partly on the testimony of the senses. Not all knowledge is empirical; some is *analytic.*

Analytic knowledge is knowledge that is arrived at strictly through the analysis of concepts.

Synthetic knowledge is non-analytic knowledge.

A priori knowledge is knowledge that is constitutive of our cognitive machinery; it's knowledge that is hardwired into us and that we therefore don't acquire.

A posteriori knowledge is knowledge that is not a priori.

Empirical truth: An empirical truth is one that can be known only through sensory observation. It is only on the basis of sense-perception that one can know that Barack Obama is currently the U.S. President. That is why "Barack Obama is currently the U.S. President" expresses an empirical truth. See "analytic truth."

Empiricism: The doctrine that all knowledge is derived from sense-perception. See "rationalism" and "sense-perception."

Commentary: Many hold that it's "scientific" to accept empiricism and "unscientific" to reject it. If, by "scientific," one means "indicative of an inability to grasp principles of any sophistication," this is correct. Otherwise, it's false. If all knowledge were observation-based, knowledge of that very fact couldn't be observation-based. In order for observation to tell you anything, it must be presupposed that observation is a source of knowledge. Therefore, observation can tell you that observation is the only source of knowledge *only* if it's known independently of observation that observation is the only source of knowledge. So, were it known that all knowledge was observation-based, it wouldn't be a truth that all knowledge was observation-based. Therefore, empiricism, if true, is either false or it cannot be known to be true. Therefore, empiricism is incoherent, and therefore false, or false. Therefore, it's false.

Encode: Anything that has a meaning is said to encode it. The content of a perception (i.e., what it tells you) is what it is "coded into it" (i.e., is what is encoded in it).

Entailment/entails: P entails Q just in case it's impossible for Q to be false if P is true. Thus *Smith has five cars* entails *Smith has more than one car*, since it isn't possible for the second to be false if the first is true.

P entails Q if P is incompatible with not-Q. See "incompatible."

Entailment (alternate definition): P entails Q just in case there is no coherently conceivable circumstance where P is true and Q is false. So *x is a triangle* entails *x has three sides* because there is no coherently conceivable circumstance where the first is true and the second is false. See "coherently conceivable circumstance."

Entailment (Second alternative definition): P entails the *negation* of Q just in case P and Q are incompatible. See "compatible."

Epistemic operator: An *epistemic* operator is one that, when joined with a sentence, yields a sentence that describes somebody's beliefs or thoughts or feelings about the proposition expressed by the original sentence. Thus, "Jerry believes" is an epistemic operator, since, when placed in front of "that 1 + 1 = 2," yields:

(iii) "Jerry believes *that 1 + 1 = 2*,"

which describes Jerry's attitude towards the proposition meant by "1 + 1 = 2." And "Laura fears that" is an epistemic operator since, when placed in front of "*that law school will be intolerably boring*" yields

(iv) "Laura fears *that law school will be intolerably boring*,"

which describes Laura's attitude towards the proposition meant by "that law school will be intolerably boring."

Any sentence containing an operator also contains an expression denoting a proposition; and, in that sentence, that operator is attributing some property to that proposition. (iii) contains the operator "it is possible that." It also contains "that 1 + 1 = 2," which denotes a proposition. And (iii) ascribes a property (that of being possible) to that proposition. See "modal operator"

and “operator.”

Equivalence: P and Q are equivalent if each entails the other. This means that there is no coherently conceivable situation where the one holds but the other does not. So *1 + 1 = 2* and *triangles have three sides* are equivalent. Note: two statements that are equivalent may have little or nothing to do with each other in terms of content. We’ll revisit this point later. See “equivalent.”

Equivalent: Two statements are equivalent if they entail each other. P and Q are equivalent iff P↔Q

Commentary: Equivalent statements cannot differ in *truth-value;* in other words, they are both true or both false (one can’t be true while the other is false). But non-equivalent statements can have the same truth-value. (“JMK is a U.S. citizen” is true, and so is “JMK is a philosopher professor.” But they’re not equivalent.).See “entails.”

“Exactly if”: A synonym of “iff” (which is short for “if and only if”) and “exactly if” and “just in case.” See “iff.”

Existential generalization (first meaning): It’s obvious that some sentences have the form “there is at least one thing that has property phi.” Here are some examples:

(i) “There is at least one thing that is a human that plays golf,”

(ii) “At least one number greater than two is an even prime.”

(iii) At least one person wrote a book called *War and Peace.*

(i)–(iii) are known as *existence-claims*, since each says that there exists a thing of a certain kind. Another term for “existence-claim” is *existential-generalization.* Existence-claims don’t always begin with “there is.” Often they begin with “something” or “someone.” Thus:

(iii*) “someone wrote a book called *War and Peace*.”

is an existence-claim.

The differences between (iii) and (iii*) aren’t relevant to what we’re doing. As far as we’re concerned, they’re the same statement.

Existential generalization (second meaning): An inference of the form “*O has phi; therefore, something or other has phi*,”—for example, “Bob is over 7-feet tall. Therefore, something is over 7-feet tall.”

Export/exportation: To “export” an expression occurring in some sentence is to move it from the inside of that sentence to the outside. Consider the sentence “snow is not white.” That is equivalent with “it is not the case that snow is white.” In the latter sentence, the negation-sign has been pushed to the outside of the sentence. For this reason, the second sentence, unlike the first, makes it clear that the ***scope*** of the negation is (in each sentence) an *entire* sentence (viz. “snow is white”). Making the meanings of sentences clear almost often involves expression-exportation.

Expression-types vs. expression-tokens: A given word can occur on several occasions. There have been countless utterances of the word “snow.” An utterance (or inscription) of that word is called a *token* of that word. The word itself is called a “word-type.” So word-tokens are bits of noise, ink, etc. Word-types are things of which these noises (etc.) are instances.

“Exactly if”: See “iff.”

Epistemically possible: A statement S is epistemically possible iff, given *only* what you know, it isn’t ruled out. So if, given only the data at your disposal, it cannot be ruled out that Smith is not in Greenland, then “Smith is not in

Greenland" is an epistemic possibility. It's obvious that epistemic possibility is a relative notion. At a given time, a given statement may be epistemically possible to some and not epistemically possible to others; and a given statement may be epistemically possible to a given person at one time but not at some other time. (If you go to Greenland, and see that Smith is there, then it's no longer epistemically possible—for you—that Smith is not in Greenland.)

Equivalent: Two statements are equivalent if each entails the other. See "entails."

Form of an expression: The sentences:

(i) "Sally hates Bob"

and

(ii) "Bob adores Amanda"

have the form

(iii) x bears relation R to y.

This is because (i) is what results when the variables in (iii) are replaced with constants, and so is (ii). In general, for two sentences S_1 and S_2 to have the same form is for there to be some open-sentence S_3 such that each of S_1 and S_2 is what results when the free variables in S_3 are replaced with constants.

Formal entailment: See "formal truth."

Formal truth: Formal truth is a property of sentences.[443] A sentence is formally true if every sentence having the same form is true. So "if Bill is tall, then it is not the case that it is not the case that Bill is tall" is formally true, since every sentence of the form "if S, then not S" is true. That sentence is also an example of a **formal entailment.** S_1 formally entails S_2 if the sentence "if S_1, then S_2" is formally true.

Formalization: To formalize a discipline is to identify a set of open-

sentences such that every proposition belonging to that discipline is either an instance of one of those open-sentences or is a formal consequence of some interpretation of those open-sentences.

Free variable: See "open-sentences."

Function: A rule that assigns no more than one object to each object in some specified class of objects. Thus, the rule that assigns the number two to the number one, and the number four to the number two, etc., is a function. (In other words, F(x) = 2x is a function.) But, importantly, not all functions involve numbers—a fact that Frege turned to good account.

Functionalism: The doctrine that for x to be a mental state of a given kind (e.g., for it to be a belief that snow is white) is for it to have certain effects and certain causes. (More formally: given any mental category M, there are certain causes and certain effects such that, given any entity x, x falls into M iff x has those causes and those effects.) Thus, anything that has the causes and effects of (say) a belief that snow is white is itself, for that very reason, a belief that snow is white. See "materialism."

Hedonism (first meaning): The psychological doctrine that nobody can do anything other than pursue his or own pleasure. One problem with this doctrine is that not all forms of well-being are identical with the experiencing of pleasure. In fact, not all forms of *enjoyment* are identical with pleasure. I enjoy playing tennis. I experience pleasure when I have a few beers. My agency is implicated in the enjoyment I derive from tennis. The joy I derive from tennis is joy that I *earn*. By contrast, the joy I derive from alcohol is not joy that I earn. My agency isn't involved. (My agency is involved in lifting the bottle to my mouth. But whereas that act is only a means to the actual source of pleasure—namely, the biochemical reactions that take place in my liver—the act of hitting a good backhand is *itself* a source of joy; it is constitutively, as opposed to merely instrumentally, involved in the joy I experience when I play tennis.) The word "pleasure" denotes enjoyment of this (non-agential) kind. The contempt that puritans have for pleasure is rooted in the fact that, whereas one's agency and, therefore, one's self are implicated in the joy that (for example) playing tennis brings, they are not implicated in the joy brought by narcotics-use and (with some qualifications)

by sexual activity. So even though it probably isn't psychologically healthy to have such a severe attitude towards such pleasures, the puritan's view of them embodies the important insight that, so far as one seeks pleasure, one seeks an effacement of self. Thus, a certain respect for human beings (*qua* agents) is inherent in the puritan's view, and a respect for people (*qua* agents) is *absent* from the hedonist's view. Where the puritan goes wrong is in his failure to see that a psychobiological prerequisite for mental health, and thus for retention of one's agency, is a certain amount of passive gratification.

Hedonism (second meaning): The ethical doctrine that nobody ought to do anything other than pursue his or own pleasure.

Higher-order property: A property of a property. The property of being a bird has many properties. It has the property of being such that there are instances of it. In other words, it has the property of being *instantiated.*

The property of being a golden mountain does *not* have that property, since there are no instances of it. It thus has the property of being *uninstantiated.*

Instantiatedness and un-instantiatedness are the two most important higher-order properties.

The property of being a person has the property of being such that each of its instances also has the property of being a mammal.

Many sentences (propositions) that appear to be about individuals are in fact about properties. Frege discovered this, and it is undoubtedly the most important insight in the history of semantics. The expressions "exist" and "doesn't exist" refer to higher-order properties. So "even numbers exist" means: *the property of being an even number is instantiated.* Thus, when appropriately re-parsed, "exist" proves to mean *is instantiated* and "even number," in that context, proves to mean *the property of being an even number.*

And "even prime numbers greater than two don't exist" means: the property of being an even number greater than two is un-instantiated. Thus, when that sentence is duly re-parsed, "don't exist" means *is un-instantiated*, and (in that context) "even prime numbers greater than two" means: *the property of being an even number greater than two.*

"Iff," "just in case," "exactly if," "↔': "Iff" abbreviates "if and only if." "P iff Q" means that P and Q are equivalent. "just in case" and "exactly if" are synonyms of "iff." So "P just in case Q" means the same thing as "P iff Q," and each means the same thing as "P exactly if Q"; and each of those means the same thing as "P→Q."

Incompatibilism: The doctrine that there cannot be personal freedom in a deterministic world. "Compatibilism" is the doctrine that incompatibilism is false. See "determinism" and "compatibilism."

Incompatible: P and Q are incompatible if they can't both be true. Thus, P

and Q are incompatible if, supposing that the one is true, the other is false. So *JM is a philosophy professor* is incompatible with *JM is not an instructor of any kind*, since they cannot both be true. See "compatible."

Identity of Indiscernibles: If x has every property that y has, and vice versa, then x is numerically identical with y.

Indeterminism: The doctrine that some things are uncaused. See "determinism."

Indexical: An indexical is an expression whose referent depends in a systematic manner on the context of utterance. (More precisely, an indexical is an expression-type E such that the semantic rule for E is to the effect that what a given token of E refers to is a function of some fact about the context of utterance.) If, on April 25, 2009, I say "today it's sunny," my utterance of "today" picks out April 25, 2009. If, on Feb. 4, 2024, I say "today it's not sunny," my utterance of "today" picks out Feb. 4, 2024. (That is why those utterances don't contradict each other and, therefore, can both be true.) In general, an utterance of "today" refers to the day on which that utterance occurs. Thus, what such an utterance refers to depends in a principled (rule-governed) and therefore systematic way on some fact about the context. Thus, "today" is an *indexical.* Other examples of indexicals are: "yesterday," "now," and "here." See "ambiguity."

Indicative conditional: A conditional that isn't a counterfactual conditional. In other words, the antecedent of an indicative conditional isn't assumed to be false (though it isn't assumed to be true). "If Smith is at the airport, then he's probably at the baggage carousel" is an indicative conditional, since it is left open whether or not Smith is at the airport. "If Smith *had* been at the airport, he *would* have been at the baggage carousel" is a counterfactual conditional, since it is being presumed that Smith was *not* at the airport. See "conditional" and "counterfactual."

Indiscernibility of Identicals: Same as Leibniz's Law (if x has every property that y has, and *vice versa*, x and y are numerically identical). See "numerical identity."

Individual: In analytic philosophy, the word "individual" has two meanings.

First meaning: An individual is anything *about which* statements are being made. Rocks, trees, and lizards are individuals in a given context, provided that, in that context, statements are being made about them. If, in that context, no statements are being made *about* the property of being green, then, in that context, that property doesn't qualify as an "individual." But in other contexts it might so qualify. Second meaning: Individuals are discrete, spatiotemporal entities—they are "things," in the narrowest sense of the word. For reasons that are given in Chapter 16, this means that an individual is any ***causal series.*** See "causal series."

Informal analytic truth: A statement is analytically true if it's non-empirically true (i.e., if its negation is prohibited from being true by the structures of concepts and by logical laws). An analytic truth is informal if other statements of the same form are false. Thus, *any case of knowledge is a case of belief* is an analytic truth, since its negation is incoherent, and it's not formally true since it has the same form as *any case of knowledge is a case of happiness*, which is false. See "analytic" and "entailment."

Instance of an open-sentence: Given some open-sentence O, if some actual sentence S is what results when the free variables in O are replaced with constants, S is an *instance* of O. So "Smith is tall" is an instance of ‹x is tall›, and "Smith is tall" is also an instance of ‹Smith has property phi.› See "open-sentence" and "quasi-quotation."

Instance of a property: If something has a property, then it is an *instance* of that property. We are all *instances* of the property of being human. Instances of properties are not identical with those properties themselves. I am not identical with the property of being human since my fellow humans are not instances of me. See "instantiate" and "property."

Instantiate: Anything that has a given property is an instance of it; and anything that is an instance of a given property *instantiates* it. You are human; so you instantiate the property of being human. See "property" and "instance."

Intensional vs. extensional definition: If a class has infinitely many members, then it must be defined *intensionally.* An *intensional definition* of a class is one that picks out the members of that class by citing some property

had by all and only those members. So the class of even numbers is *intensionally defined* as the class that contains every number that is divisible by two and that doesn't contain anything else.

If a class has only finitely many members, then it is possible to give an *enumerative* definition of it. Consider the class of even numbers greater than zero but less than 10. It can be enumeratively defined as the class that contains 2, 4, 6, and 8.

Technically, enumerative definitions are intensional definitions. If you identify a class by listing its members, you *are* identifying a property that they and they alone have. For example, if you identity C as the class containing 2, 4, 6, and 8, you are identifying C as the class containing all and only those objects x such that x have the property of being identical with 2 or 4 or 6 or 8. See "intension vs. extension."

Intension vs. extension: Intensions are to predicates what senses are to referring terms, and extensions are to predicates what referents are to referring terms. So if Beethoven's nine symphonies are Smith's nine favorite pieces of music, then "symphony composed by Beethoven" has the same extension as "one of Smith's top nine favorite pieces of music," meaning that, for any object x, ‹x is one of Smith's top nine pieces of music› is true iff ‹x is a symphony composed by Beethoven.›

But obviously those sentences have different intensions. After all ‹x is a symphony composed by Beethoven› entails ‹x is a symphony›, whereas ‹x is one of Smith's top nine favorite pieces of music› does not entail that. See "predicate" and "quasi-quotation marks."

Interpretation: As it is used in everyday life, an "interpretation" of a given body of data is a hypothesis as to what is meant by that data. As it used in philosophy, the word "interpretation" sometimes bears this meaning; but it sometimes bears a very different one. An "interpretation," in the distinctly philosophical sense of the word, is an *assignment* of meaning to an expression that contains undefined or only partially defined constituents. (In what follows, the words "interpretation" and "interpret" (etc.) will be used only in the second, strictly philosophical sense.)

It is statement-forms, not statements proper, that are interpreted. (A statement-form is an ***open-sentence.***) The reason is that anything with an undefined constituent *ipso facto* fails to say anything true or false. To interpret an open-sentence is to assign constants to the variables. So one way to interpret "x is even and y is odd" would be to assign the numbers two and three to "x" and "y," respectively. That interpretation of that statement-form *validates* it, meaning that the resulting sentence is correct. By assigning three and two to "x" and "y" respectively, we've produced an interpretation of it that fails to validate it. And by assigning Barack Obama to "x'" and the property of being the U.S. President in 2009 to "phi," I've produced an interpretation of "x is phi" that validates it.

Oftentimes, expressions that *seem* not to contain free variables do. For example, "Jim is a nice guy" is so vague, that until a specific meaning has been assigned to "nice," nothing has really been said. Where expressions of this sort are concerned, context tends to supply the missing interpretations, that being why we don't see "nice" as being a variable, like "phi." Also, with "nice," there is an understood, and quite restricted, range of possible interpretations, whereas with "phi," that range is likely to be more open-ended. Thus, we're quick to see "phi" for the variable that it is, but not so quick to do so with "nice."

"Just in case": See "iff."

Leibniz's Law: "According to *Leibniz's Law*, if x and y are numerically identical, then the one has any property had by the other. So if Ben Franklin is smart, so is the inventor of bifocals."

Logical form: The logical form of a sentence S is what is represented by a

sentence S* that makes it clear what S entails and what entails S. Consider the sentence:

(a) “nothing snores.”

Given its grammar, (a) seems to license the inference that there exists some strange entity—the nothing, or some such—that snores. But (a) doesn’t warrant this inference. What (a) means is perspicuously represented by:

(b) “the property of snoring is un-instantiated.”

Unlike (a), (b) isn’t misleading as to what can be inferred from it. But (a) and (b) encode the very same proposition. Since (a) and (b) satisfy these conditions, (b) is a representation of (a)’s logical form. If the relationship borne by a sentence S_1 to another sentence S_2 parallels that borne by (a) to (b), S_2 gives S_1’s logical form. See “entailment” and “meaning vs. entailment-relations.”[444] See “logically perfect language.”

Logically perfect language: Let L be a language having the following property. Given any sentence S belonging to L, S’s logical and grammatical forms coincide. In that case, L is a *logically perfect language.*

There can be no such language. There could exist such a language only if a sentence’s entailment-relations could always be read off of its grammatical surface. This would be possible only if all entailment were *formal* entailment. But entailment is not, in general, formal entailment. Formal entailment is parasitic on informal entailment, and formal analytic truth is parasitic on informal analytic truth. The reasons for this are given in Chapters 1 and 18.

Main connective: See “connective.”

Materialism: The doctrine that mental entities (e.g., beliefs) are identical with physical entities (e.g., brain states).

Meanings vs. entailment relations: For two sentences to have the same meanings, it is *necessary* that they be equivalent (i.e., that they be such that the one is true if and only if the other is true). So if P entails something that Q does not entail, then P and Q don’t have the same meaning. For example: “Jim is an avid reader” entails that Jim can read, whereas “Jim is a lousy

soccer player" does not entail that. Therefore, those two sentences have different meanings.

But even though, in order for two sentences to be equivalent, it is *necessary* that the one entail any sentence entailed by the other, that isn't *sufficient* for sameness of meaning. For any x, ‹x is a unique even prime› is equivalent with ‹x is the unique number n such that n = 2 if there is no completeness proof for arithmetic and n = 3 otherwise.› Neither sentence entails anything not also entailed by the other. But those sentences obviously differ in meaning.

For two sentences to have the same meaning, it is necessary, not only that they entail the same things, but that they do so *in the same way.* Even though, for any x, ‹x is a unique even prime› entails the same things as ‹x is the unique number n such that n = 2 if there is no completeness proof for arithmetic and n = 3 otherwise›, they entail those same things in different ways. For example, the *way* that ‹x is even› is derived from the first is very simple, whereas the way it is derived from the second is very complex.

Modal operator: A modal operator is an ***operator*** (i.e., a device that, when coupled with a sentence or ordered n-tuple of sentences, yields a new sentence) that describes the *modal status* of that proposition, meaning that it says of that proposition that it is can be true, must be true, can be false, or must be false. Thus, "it is necessary" is a modal operator, since, when placed in front of "that
1 + 1 = 2," yields:

(i) "it is necessarily the case that 1 + 1 = 2,"

which describes the proposition meant by "1 + 1 = 2" as necessarily true. And "it is possible" is a modal operator, since, when placed in front of "that Smith is in France," yields:

(ii) "it is possible that Smith is in France,"

which describes the proposition meant by "Smith is in France" as potentially true.

In (i), "that 1 + 1 = 2" is said to be *governed* by "it is necessarily the case," meaning only that it denotes the proposition to which the property of

being necessarily true is being attributed. In (iv), "that law school will be intolerably boring" is governed by "Laura fears" since it denotes the proposition to which the property of being feared by Laura is being attributed.

Anything within an expression governed by an operator is said to fall within the *scope* of that operator. In (ii), "France" falls within the scope of "it is possible." In (iii), "two" falls within the scope of "Jerry believes that." See "operator" and "epistemic operator."

Model: To "model" a collection of data is to produce a hypothesis that accounts for it.

Model (second meaning): A model of a set of open-sentences is an assignment of meanings to the undefined, or partially defined, expressions occurring in a set of open-sentences. Consider the following open-sentences:

(1) There is nothing that bears R to N.

(2) Anything x bears R to one, and only one, thing y.

(3) If x bears R to y, and y bears R to z, then x bears R to z.

(4) If x bears R to y, y doesn't bear R to x.

Whether (1)–(3) come out true depends on what we take "N" and "R" to stand for, and also on what we take "anything" and "nothing" to mean. Do we mean *anything at all?* Or do we mean anything *falling* into some specific class of objects (e.g., the class of numbers or of aquatic mammals)? The same question *mutatis mutandis* arises in connection with "nothing." (If we take "anything" to mean "any whole number," then the *universe of discourse* is the class of whole numbers. If we take it to mean "any penguin," then the universe of discourse is the class of penguins.)

To answer these questions is to provide an "interpretation" of these expressions. An "interpretation" of a set of open-sentences is thus an assignment of definite meanings to the undefined or partially defined expressions occurring in it. If, given a particular interpretation X (1)–(3) come out true, X is a "model" of (1)–(3). If, given X, (1)–(3) comes out false, then X doesn't model them.

Here's one interpretation of (1)–(3). Take the universe of discourse to be

the whole numbers (zero, one two, etc.). N is zero, and R is the relation of being the immediate predecessor of. Relative to this interpretation, (1)–(3) come out true. This interpretation is therefore a *model* of (1)–(3).

Here's a different interpretation of (1)–(3).Take the universe of discourse to be the class of people. Take N to be Dick Cheney, and take R to be the relationship that a child bears to either one of its parents. This interpretation fails to model (1)–(3). There are several reasons for this. Thus, interpreted, (1) says Dick Cheney isn't a parent, which is false; (2) says that nobody has more than one parent, which is false; and (3) says that grandparents are immediate parents, which is false.

Molecular proposition: A proposition that has another proposition as a proper part. For example, *if grass is green, then something is green* is a molecular proposition, since it has two propositions as proper parts of itself, namely, *grass is green* and *something is green.* Other examples of molecular propositions are *Bob is short and Mary is tall; Mary is smarter but Sally is smarter than Mary.*

Some propositions that are in fact molecular don't *appear* to be at first. For example, *John wants to catch a fish* is a molecular proposition, since it's the same as *John wants it to be the case that John catches a fish,* and *John catches a fish* is a proposition. See "atomic proposition."

Molecular sentence: A sentence that has another sentence as a proper part (e.g., "grass is green and snow is white"). Some sentences are molecular that don't initially appear to be. For example, "John wants to catch a fish" is molecular since it's really an abbreviation for: "John wants it to be the case that John catches a fish."

Many of the expressions we're about to define (e.g., "quantified generalization") are used in connection with propositions *and* with sentences.

Mutatis mutandis: This means *provided that the relevant changes are made.* Suppose that you and I have different employers and also that your boss hates my boss with such vitriol that he will reward anyone who harms my boss in any way. In that case, the statement: "I'll get fired if I punch my boss, and you'll get fired if you do the very same thing" is false. Supposing that I punch my boss, your doing the very *same* thing would consist in your

punching *my* boss, which wouldn't get you fired. But if you were to punch *your* boss, you would get fired. Thus, the right statement is: "I'll get fired if I punch my boss, and you'll get fired if you do the very same thing *mutatis mutandis.*"

Necessary: A proposition is necessary if there is no coherently conceivable circumstance where it is false. Example of such a proposition: 1 + 1 = 2. See "contingent," "possible," and "proposition."

Negation: The negation of P is not-P. So the negation of a proposition is the statement that says that P is false. (Ockham uses the obsolete term "contradictory opposite" instead of "negation.")

Necessity, sufficiency: "If P then Q" means that there's no way that P can be true *unless* Q is true. This, in its turn, means that the truth of Q is *necessary* for that of P and, also, that the truth of P is *sufficient* for the truth of Q. See "entails."

Necessity, sufficiency (revisited): "P entails Q" means the same thing as "Q is necessary for P," which in turn means the same thing as "P is sufficient for Q." See "entails."

Normative proposition: A proposition is *normative* if it says how things *should be. Killing is wrong* is a *normative* proposition, since it says that killing *shouldn't* occur. See "proposition" and "descriptive proposition."

Numerically identical: x and y are numerically identical if they are the very same thing. So Benjamin Franklin is numerically identical with the person who invented bifocals. If you punch Ben Franklin, you are also punching the inventor of bifocals and *vice versa.*

Observable characteristic: Any characteristic of a thing that one can see or otherwise sense-perceive. One can *see* that a thing is red. One can *feel* that a thing is solid. Thus, redness and solidity are, in at least some instances, observable characteristics. One cannot *see* that a thing consists of such and such micro-particles. Thus, the property of being composed of such things is not an observable one.

It's obvious that whether a given characteristic is observable or not is

typically context-dependent. A thing's being red or solid might be as theoretical a fact as its consisting of molecules of a certain kind. To creatures of our intelligence level, but of one-trillionth our size, the fact that my desk is solid could be known only through elaborate inferences; and its being solid would not, in that context, be something that could be observed.

Still, the distinction between observable and unobservable characteristics is a good one. But, when saying that x's having phi is (un)observable, one must realize that what one is saying is elliptical for a relational statement of the form: given the nature of physical relationship to x, x's having phi is, with respect to us, an (un)observable fact.

Occurrence: *An occurrence of an expression is a token of it. The expression "snow" occurs three times (i.e., there are three occurrences of it) in the following sentence: "Jim likes snow and I love snow but Mary hates snow." See "expression-token."*

Open-sentences: Consider the sentence "Bill Gates is wealthy." That sentence is true. Replace "Bill Gates" with a variable—with an "x." The result is ‹x is wealthy›, which is neither true nor false. ‹x is wealthy› is an "open-sentence." (The characters flanking that expression are known as ***quasi-quotation marks.*** They may, for the time being, be treated as quotation-marks.)

In general, an open-sentence is an expression that contains a free variable and is therefore neither true nor false.

A *free variable* is one for which no reference is supplied either by the context or by preceding material. So if, out of the blue, I say ‹x is even›, my utterance contains a free variable, and is therefore neither true nor false.

A *bound variable* is a variable that isn't free. Thus, a bound variable is one that is assigned a referent by the context or by preceding material. Consider the sentence:

(LU) "*For any human being x, if x isn't loved by anyone, x isn't happy.*"

LU says that, *given* a human being, if *that* human being isn't loved, then *that same human being* isn't happy. Thus, the italicized part assigns a referent to the occurrences of the variable (the "x"). The occurrences of "x" in the underlined part of LU are *bound.*

An open-sentence isn't true or false. But replacing the free variables in an open-sentence may result in a true sentence. If an open-sentence is such that replacing the variables in it sometimes/always/never results in a true sentence, that open-sentence is said to be "true for some/all/no values of its variables."

Thus, "x is an even number" is true for some values of its variables (since "two is an even number" is true but "three is an even number is false"); and "x is identical with x" is true for all values of its variables; and "x is not identical with x" is false for all values of its variables.

Commentary: Logicians generally say that their discipline studies relations holding among sentences. This isn't true. It studies relations holding among the sentence-schemata. "Grass is green and show is white" entails "snow is white." The logician is interested in this fact *only* to the extent that it alerts him to the corresponding generalization, viz. for *any* sentences P and Q, ‹P and Q› entails ‹P. › But ‹P and Q› and ‹P› are not sentences; they're open-sentences; they're sentence-like entities that contain free variables and, unlike real sentences, are therefore neither true nor false.

Operator: An "operator" is any expression that, when joined with a sentence, forms a new sentence. So "it is possible" is an operator, since, when conjoined with "Smith is in France," a new sentence is formed, namely "it is possible that Smith is in France." And "George believes" is also an operator for the same reason.

There are many kinds of operators; but two kinds are of special importance: modal operators and epistemic operators. See "predicates" and "quasi-quotation marks."

"Other things being equal": See "*ceteris paribus.*"

Perception: See "sense-perception."

Performative utterance: When, in the context of a marriage ceremony, the priest says "I now pronounce you man and wife," the priest is not (so it would seem, and so we'll momentarily assume) reporting an existing fact. (I'll explain the raison d'être for the hedge in a moment.) He is not affirming the already existing truth that he is now in the process of forging a marital bond between the two individuals before him. In uttering those words, he is

creating a fact, not reporting a fact.

A sentence-utterance functioning in this way is known as a "performative" utterance. Other examples of Performatives are utterances of "I christen this ship the Mel Gibson," "I hereby challenge you to a duel," and "I promise to pay you back with interest if you loan me the money."

The concept of a performative was first clearly identified by J.L. Austin in his (1955) lectures at Harvard, which were published in 1976 under the title *How to do things with Words.* Austin's points are anticipated, albeit obscurely, by Wittgenstein (1958).

Austin said, very plausibly, that performative utterances are neither true nor false. In saying "I now pronounce you man and wife," the priest isn't describing an existing fact, and his words are therefore neither true nor false —or so it would seem.

But in his (2004) book *Philosophy in the 20th Century* (Volume II), Scott Soames brilliantly shows that this is false. Consider the statement:

(S) "Supposing for argument's sake that I now pronounce you man and wife, it is incumbent on you to love and cherish each other."

(S) is synonymous with:

(S*) Supposing it true that I now pronounce you man and wife, it follows that it's true that it is incumbent on you to love and cherish each other."

S is meaningful; in fact, it's correct (or at least could be, given certain reasonable assumptions about the individuals in question). Let O be the occurrence in S of "I now pronounce you man and wife." If O cannot be meaningfully supposed true, S is meaningless, and therefore neither true nor false. Bearing this in mind, let O* be some occurrence of "I now pronounce you man and wife" that is occurring on its own. O and O* obviously have the same meaning. Since O is capable of being meaningfully supposed true, the same is true of O*, and Austin is therefore wrong to say otherwise.

An argument similar to the one just given shows that a celebrated contention of Wittgenstein's is false. Wittgenstein (1958) said that, when one says "I am sad"—when, in general, one makes any statements about one's state of mind —one isn't really *saying* anything. Saying "I am sad," Wittgenstein says, is no more a *bona fide* speech act than is grimacing or moping about. In saying

this, Wittgenstein was trying to undermine the view that sentences report facts; he was trying to vindicate his provocative—but, when scrutinized, quite hollow—contention that "words are deeds." (This contention is one to which Austin gave some substance in his theory of performative utterances.)

We can use an analogue of Soames' argument to refute this contention of Wittgenstein's. If, as Wittgenstein alleges, a stand-alone occurrence of "I am sad" is neither true nor false, as it's merely a glorified frown, then it's counterpart in an utterance of

(S#) "if it's supposed true that I am sad, then I am in a mood that is, in at least some respects, less than optimal"

is equally meaningless, an immediate corollary being that any utterance of S# is meaningless. But such an utterance is quite meaningful; in fact, such utterances are correct.

Also, if you say "I am sad," you are obviously reporting a fact. That's why your utterance could be false—you could be lying after all. Thus, utterances of "I am sad," and first-person reports generally, can be meaningful, and usually are, contrary to what Wittgenstein is alleging.

Personal stratum of cognition: The totality of mental events within one's own mind that could in principle fall within the scope of introspective awareness. Mental events in your mind mediate between sensory input and cognitive output. You could not become *introspectively* aware of these. Nor are you aware of the cognitive events in your own mind that mediate language-learning and comprehension. So far as you can become aware of them, it is in the way that you can be aware of sub-atomic particles, and not in the direct, introspective way in which you are aware of your own sensations. No amount of psychoanalytic intervention could make you introspectively aware of them, and they are therefore "deeply" unconscious, to use Chomsky's term. But, given psychoanalytic intervention, you could become introspectively aware of the unconscious events and conditions posited by Freud. Thus, the Freudian unconscious falls within the scope of the *personal* stratum of cognition, as does everything that, without such intervention, is accessible to introspection. The cognitive processes mediating language-comprehension, perception, etc., are not done by *you*, even though they happen *in* your mind. There is thus a sense in which they are impersonal

(or subpersonal). They constitute a framework underlying the events constitutive of the personal stratum of cognition and, for that reason, are accessible to it.

So far as I know, Chomsky (1959) deserves credit for being the first to delineate the concept of subpersonal thought, and also for being the first to put the concept of subpersonal mentation to scientific use. In doing the former, he opened up new vistas, not just in psychology, but in epistemology and philosophy generally. There are allusions to the subpersonal in Chapter VII of Freuds (1901) work *The Interpretation of Dreams.* The main contention of Kant's (1789) work, *The Critique of Pure Reason*, presupposes its existence. Perception and thought, Kant contends, result from a convergence of (i) disturbances of our sensory surfaces and (ii) concepts that are hardwired into us and constitute our innate cognitive endowment. Kant makes it very clear that it is not until *after* these concepts have processed these disturbances that there occurs any mental activity that could be attributed to a *person*. So Kant clearly believes that personal mental activity presupposes the existence of pre-personal—or, as we might also put it, subpersonal—mental activity. Chomsky often cites Kant as one of his forbears.

Perspicuous/perspicuity: A sentence is "perspicuous" if it is clear *in the sense that* its inferential structure (what it entails and what entails it) can be read off of its grammatical structure. A perspicuous, set-theoretic translation of "2 + 2 = 4" would be totally unclear to most people, even though "2 + 2 = 4" is very clear (in the psychological sense). That's why "clear" and "perspicuous" aren't interchangeable. Perspicuity is a logical property; clarity is a psychological property. Perspicuity is to some extent a contextual notion. Depending on what one's background assumptions are, one sentence may or may not be more perspicuous than some other synonymous sentence. But, other things being equal, the more syntactic structure a sentence has, the more perspicuous it is. The reasons for this are given in Chapter 18, Section 4.1.)

Possible: A proposition is possible if there is some coherently conceivable circumstance where it is true. Example of such a proposition: Mary owns a BMW. See "contingent," "necessary," and "proposition."

Predicate: In the sentence, "Smith is tall," the *predicate* is "tall." A "predicate" is generally said to be an expression that refers to a property. But this isn't a good definition. The expression "the property of being tall" is not a predicate; it's actually a singular term. But "tall" *is* a predicate. So is "smart."

In general, a predicate is an expression that, when joined in a grammatically acceptable manner with a singular term, yields a sentence that attributes a property to the individual referred to by that term. "Smart" is a predicate because, when joined (in a grammatically acceptable way) with "Sally," the result is a sentence that attributes smartness to Sally, viz. "Sally is smart."

Predicates may be identified with open-sentences. See "intensions vs. extensions" and "open-sentences."

Primitive expression: Synonymous with "semantically simple expression." A primitive expression is one that doesn't consist of other expressions. Examples of such expressions are "red" or "sweet."

Occurrences of primitive expressions are not as easy to identify as one might think. Let T be an utterance of "snow is white." One would think that, in T, the occurrence of "snow" is semantically simple.

Chomsky would say that it contains a phonetically unrealized case-marker. Chomsky's reasons for taking this view are extremely compelling. See Chomsky (1965, 1998).

Probability: This word is "ambiguous." It has (at least) two meanings. First, there is the statistical meaning of "probability." If there are 10 balls in the urn, only one of which is white, the statement "x is a white ball in the urn" has a statistical probability of 1/10.

Then there is explanatory probability. P makes Q probable in the *explanatory* sense if, given what it is that is already believed, there are more causal anomalies if P is true and Q is false than there are if, other things being equal, P is true and Q is true.

It is argued in Chapter 11 that explanatory probability cannot be reduced to, or understood in terms of, statistical probability.

Property: A property is a characteristic. Other words for "property" are

"attribute" and "feature." This is the same as saying that a property is anything of which there can be instances. Put another way, a property is anything that can be *had*—"had" in the sense in which one has a characteristic, not in the sense in which one has a car. So the "had" in question is that of *attribution*, not of *possession.* A property is anything that can be meaningfully attributed to something.

Commentary: some exotic properties: Consider the number two. There are *instances* of this property. Any pair of objects is such an instance. The number two can thus be thought of as a property had by all pairs of objects and by nothing else.

Anything of which there are *instances* is a property. (But there are some properties of which there are no instances; e.g., the property of being a perfectly moral human. But that property is composed entirely of things of which there are instances. So it's possible—though I haven't personally verified it—that every property *either* has instances *or* is composed of properties that have instances. Question: are there properties that counterexample this conjecture?)

The *numeral* "2"—that is, the *sign* for the number two—is *also* a universal. But it's distinct from the number two itself. The numeral "2" is, for reasons that we discussed a little while ago, a property of certain ink marks and bursts of sounds. (Those ink marks, etc., are *tokens* of that numeral.) But the number two is a property of *pairs* of objects. "2" is no more identical with the number two than the name "John-Michael" is identical with the person John-Michael.

Works of music are properties. *Instances* of those properties are what we refer to as "performances." Consider the first movement of Beethoven's "Moonlight Sonata." I played it a few minutes ago. But that work of music didn't cease to exist when I stopped performing it. Of course, it's likely that other people are performing it. But it wouldn't cease to exist if, for a 10-minute period, *nobody* were to perform it. The "Moonlight Sonata" isn't identical with any *particular* stream of sound. Therefore, it isn't identical with any stream of sound at all. The "Moonlight Sonata" is a *property* of certain streams of sound—of streams of sound that have certain melodic and harmonic properties. When somebody plays that Sonata on the piano, that

person is producing a stream of sound that has the right properties and is therefore an instance—or, as we would usually say, a performance—of that sonata. (Other works of art, e.g., paintings, appear to be concrete objects. The Mona Lisa is a concrete object; it could be destroyed. Works of visual art seem, in general, to be spatiotemporal and thus to be property-instances, as opposed to properties. Whether works of literature are properties or property-instances is a delicate question that will be set aside.)

Propositions: The things that are *meant* by true or false utterances or inscriptions.

Commentary: Propositions aren't sentences; propositions are sentence-meanings, not sentences per se.

Propositions are not spatiotemporal. The proposition *snow is white* does not have a location in space or time, even though particular *instances* of white snow obviously do have such a location.

Propositions are not psychological entities; they are mind-independent entities.

Believing a proposition is something mental; but the thing believed is not. More generally, *grasping* a proposition is something mental; but the thing grasped is not.

Let's suppose that:

(1) "Smith believes that the inventor of bifocals is 6-feet tall'

is true.

What is the *object* of Smith's belief? What is the *thing believed in*—the thing that, in having that belief, he believes to be true?

It's tempting to say that it's something *mental*—that it's some mental event or psychological condition. But this is wrong. Smith's *having* that belief is indeed a psychological condition. But the thing *believed* is not. The believ*ing* is mental; the believ*ed* is not. In other words, the thing that, by virtue of having that belief, Smith regards as true isn't a mental or psychological event or condition (or whatnot) at all. Here's why.

Whatever it is, that thing is something that *others* grasp. (And, what

follows, it is also something people besides Smith *could* grasp, even if they don't.) Those who agree with Smith (about the height of the inventor of bifocals) believe it true; those who disagree believe it false. But they all grasp that thing.

Supposing that any given one of those people were to perish, and that all his or her mental contents were therefore to vanish, these other people could (and probably would) continue to grasp that thing; they could continue to believe it true or false or probable or whatnot. If Smith and I both believe that the inventor of bifocals is 6-feet tall, Smith's dying won't prevent me from continuing to believe it or, therefore, from continuing to *be able* to believe it. Smith's death, though tragic, is irrelevant.

By the same logic, appropriately generalized, given *any* two people who believe or otherwise grasp *any* proposition, the existence of the one person's mind and mental states is independent of the other person's believing (or otherwise grasping) that proposition. *Any* one person's mental states and events can cease to exist without in any way jeopardizing anyone's ability to believe or disbelieve that the inventor of bifocals is 6-feet tall.

Thus, the thing that Smith believes—the *object* of his belief—is something that continues to exist even if any given person's mental events and conditions cease to exist. Therefore, that thing isn't identical with anyone's mental states or, needless to say, with anyone's mind in its entirety. Nor is it identical with any *thing's* mental states. Supposing that non-humans —animals orMartians or whatever—have beliefs concerning the height of the inventor of bifocals, everything we just said holds of those beliefs. Since that thing isn't identical with the mind or mental states of any entity, it can't be identical with *any* mental entity. To deny this would be to affirm the absurdity that there is a mental thing which isn't any particular mental thing.

What, more exactly, is the nature of this thing? What, in general, are propositions? (In other words, what is the nature of the things that are the objects of belief and disbelief? We don't have to worry about that right now; it's irrelevant to what we're about to say. But it is thoroughly answered in Chapter 3.)

Additional commentary: Propositions have the distinctive and crucially important property of being *non-derivatively true or false.* What does this

mean? Propositions are obviously true or false. (It's true that Finland is in Europe; it's false that it's an island in the Pacific Ocean.)

To be sure, things other than propositions are true or false (e.g., sentences and beliefs). The sentence "Finland is in Europe" is true. But that sentence is true *only by virtue of its association with the corresponding proposition.* That sentence is true only because it expresses a true proposition. So it's really the proposition that deserves to be described, without qualification, as "true"; and when we say of the sentence that it's "true," we're using the word "true" to pick out a related, but distinct, property; we're using it to pick out the property of *encoding* something that's true.

And these, it must be stressed, are very different properties. It's one thing to have the property of being true. (The proposition *Finland is in Europe* has that property. It's quite another to have the property of *encoding* something true.) The sentence "Finland is in Europe" has *that* property. But the corresponding proposition does *not* have it, since propositions, unlike sentences, don't encode anything.

What we just said about sentences is true of beliefs. My belief that Finland's in Europe is "true." But it's true *only because of its association with a true proposition.* In other words, it's "true" in the sense that it has for its content something that is true. Put another way, the thing which that belief is a belief *in* is true, and the belief is "true" only in the second-class sense that it has this association with this other thing—this other being something that, unlike the sentence, is true in a first-class, non-derivative sense.

Propositions are thus "non-derivatively true," whereas beliefs and sentences are "derivatively true." For this reason, propositions are sometimes defined as "non-derivative bearers of truth and falsity."

We thus have three different definitions of "proposition": (i) thing meant by a sentence; (ii) thing towards which one can have an attitude (e.g., belief, disbelief, fear, hope); (iii) non-derivative bearer of truth or falsity. While no one of these definitions is 100% accurate each is accurate to a high degree of approximation; and while no two of them are precisely equivalent (i.e., while no two of them pick out precisely the same class of entities), any two of them are almost equivalent (i.e., the classes picked out by any two of them overlap to a very high degree).

Propositional attitudes: Propositions are the things we affirm. They are also the things towards which belief, doubt, fear, and all other *attitudes* are directed. One can't just believe; one must believe that snow is white or that grass is green. One can't just fear; one must fear *that Mary will come home before tomorrow*, or some such.

We'll see in a moment that there are things other than propositions that we believe, doubt, etc. But, though it's therefore not entirely accurate to say that propositions are *the* things we believe, doubt, etc., it's a good point of departure; and, for the time being, we'll assume it's truth.

In believing that Sally will come home tomorrow, one has a certain propositional *attitude.* In fearing it, one has a different propositional attitude. The proposition is the same, but the attitude has changed. In wondering whether Sally will come home, one has yet another attitude towards that same proposition. Belief, doubt, wonder, etc., are thus "propositional attitudes."

Knowing what is going on in somebody's mind consists largely, though by no means entirely, in knowing what that person's propositional attitudes are. And to know this, one must know which propositions are the objects of that person's mentation; and one must also know, for any given one of those propositions, which attitude that person has towards it. To know what is going in Larry's mind, it isn't enough to know that the *object* of his thoughts is the proposition *Sally will come tomorrow;* one must also know what sort of attitude he has towards that proposition. Does he believe it? Hope that it's true? Fear that it's true? And one must have comparable knowledge of every other proposition that is the object of Larry's thought. If one doesn't have such knowledge, one is ignorant of an extremely important dimension of Larry's existence; and if one does have it, one knows much about Larry—but not everything.

Why not everything? Because not everything mental is a propositional attitude. Nothing that isn't *content-bearing* is a propositional attitude. (To say of something mental that it's "content-bearing" is to say that it bears some kind of message. So my current visual perception is content-bearing, since it tells me that there's a computer screen in front of me. My headache isn't content-bearing, since it doesn't bear any message; it's merely felt.) And many mental entities that are content-bearing aren't propositional attitudes,

the reason being that not all contents are propositions. See “propositions.”

Qua: Synonym of “in virtue of.” Larry eats a fish. He’s allergic to all food-products that contain protein X. This particular fish (though not others) contains X. So Larry falls ill. Jerry eats a fish; he isn’t allergic to fish, but the fish he eats is rotten. So he falls ill. What made Larry sick wasn’t his eating a fish; it was his eating something that had X in it. That thing happened to be a fish, but that wasn’t what was operative. Thus, it wasn’t in virtue of his eating a fish, but in virtue of his eating something containing X, that Larry became sick. Put another way, it was his eating a fish *qua* thing that contained X, not a fish *qua* fish, that made Larry sick. Similarly, it was Jerry’s eating a fish *qua* rotten thing, not a fish *qua* fish, that made him sick.

Qualitatively identical: x and y are *qualitatively* identical if they resemble each other to a maximally high degree but may nonetheless be *distinct individuals.* So if you have an identical twin—a twin who looks, thinks, talks, acts (and so on) just like you—then you are qualitatively identical with that person. But you are not *numerically* identical with that person, since you are not literally the very same individual as that other person.

Quantified generalization: Any sentence or proposition that says of some property how many instances it has. Thus, *nothing snores* is a quantified generalization since it says of the property of being a snorer that it has no instances. For the same reason *mutatis mutandis*, each of the following is a quantified generalization: *everything snores, something snores, three things snore*, and *most things snore.* See “quantifier.”

Quantifier: An expression that, when coupled with a predicate yields a sentence saying how many things of a given kind have that property. A predicate may *roughly* be said to be an expression denoting a property (e.g., “snore,” “is tall”). (Why only roughly? See below.) “All people” is a quantifier, since, when coupled with “snore,” which denotes the property of snoring, yields “all people snore,” which says how many things of a certain kind have the property of snoring. “No fish” is a quantifier since, given the expression snore, yields the sentence “no fish read Shakespeare,” which says how many things of a given kind have the property of reading Shakespeare.

In logic and the philosophy of language, one often uses artificial

analogues of expressions like “all,” “some,” “none,” etc. In logic, the sentence “all people snore” is represented as “for any x, if x is a person, then x snores.” And “some people snore” becomes “for some x, x is a person and x snores.” The reason why artificial quantifiers (e.g., “for any x,” “for some x”) are used in lieu of their natural-language counterparts (“all,” “some”) is that, for the reasons discussed in the previous chapter, the logical form of quantified statements is concealed unless these replacements are made. For example, “no person snores” has the same grammatical form as “John snores,” which misleadingly suggests that, like the latter, the former attributes the property of snoring to some entity. But “nothing snores” does *not* do that; that’s the very opposite of what it does. What “no person snores” says is this: “for any x, if x is a person, then x does not snore.” The latter sentence clearly *doesn’t* attribute the property of snoring to anyone or anything, and is, in this respect, more perspicuous than its natural language counterpart.

The reason it’s only roughly correct to say that “predicates” denote properties is that “the property of snoring” and “the property of being tall” clearly pick out properties—but they aren’t predicates. See “predicate.”

Commentary: All quantifiers are considered to be connectives. Thus “for all x,” “for no x,” etc. are connectives. Here is a very rough explanation of the rationale for this practice. Consider the sentence “all people snore.” This can be thought of as saying that each possible sentence of the form “x snores,” where x is some person, is true. So it can be thought of as saying that “John snores,” “Sally snores,” are all true. Thus, “for all x,” can be thought of as an expression that, when given “Sally snores,” “John snores,” etc., as inputs, yields another sentence, viz. “it is true that John snores, Sally snores [etc.].” I must point out that this explanation is very approximate. But the technicalities that would have to be gone through to make it completely accurate are not, at this stage, very important.

Given that all quantifiers are connectives, it follows that all quantified generalizations are considered to be molecular propositions. See “quantified generalization.”

Quasi-quotation marks: In logic and the philosophy of language we often wish to talk about whole *classes* of sentences having a given form. We might wish to say, for example, that *all* sentences of the same form as “grass is green or it is not the case that grass is green” are true, that *all* sentences of the

same form as “grass is green and it is not the case that grass is green” are false, that *all* things have the property of being self-identical, that *no* thing has the property of a being a square circle.

But it’s hard to do this without producing a nonsense-statement. Consider the following statement:

(1) For any proposition P, the expression “P or not-P” is true.

(1) seems okay at first. But it isn’t. It’s false. This is because “P or not-P” isn’t a sentence. If, out of the blue, I say to you “x is larger than four,” I haven’t said anything, since “x” is a variable, not a numeral. (If I don’t say it out of the blue, and it’s understood that “x” has a certain referent, then it may be true.) Similarly, so “P or not-P” is no meaningful statement can contain a free variable. If I say to you, “P or not-P,” I haven’t said anything, the reason being that “P” is a variable. “P or not-P” isn’t true or false, and it’s no more capable of being true or false than an out-of-the-blue utterance of “x is larger four.” Given that “P or not-P” is never true, it follows that (1) is false. After all, (1) says that “P or not-P” is *always* true—when, in fact, it’s *never* true (or false).

But what (1) is *meant* to express, even though it fails to do so, is that, given the expression “P or not-P,” the result of replacing both occurrences of “P” with a true or false sentence is itself a truth. (For example, if we replace those occurrences with “snow is white,” the result is “either snow is white or it is not the case that snow is white,” which is true.)

The sentence:

(2) For any proposition P, the expression ‹P or not-P› is true

is an abbreviation for

(2) Given the expression “P or not-P,” the result of replacing both occurrences of “P” with a true or false sentences is itself a truth.

(1) is often abbreviated into:

(3) For any P, ‹P or not-P› is true,

which, in its turn, is often abbreviated into:

(4) P or not-P.

The strange rectangular expressions in (3) are known as “quasi-quotation marks.” Putting quasi-quotation marks around an expression results in an expression that refers to everything in the original expression *besides the variables.* So ‹if x is larger than y and y is larger than z, then x is larger than z› is an expression that refers to everything in the expression in between the quasi-quotation marks *besides* the occurrences of “x,” “y,” and “z.” Quasi-quotation marks are the invention of Willard van Orman Quine (1941).

Quasi-quotation marks are extremely useful. Without them, or some other comparable device, we couldn’t express the truth expressed by the sentence:

(4) For all values of x, y, and z, the sentence ‹if x is larger than y and y is larger than z, then x is larger than z› is true.

Quasi-quotes, though seldom *overtly* present in discourse, are very often *implicitly* so. I've shown elsewhere that, without them, it isn't possible to make what is meant by expressions that have "discourse-internal referents,"—for example, the occurrences of "that person" in the sentence: "if a person is mean to everyone, that person will have no friends and that person will therefore be unhappy."

Some more illustrations: The sentence:

(5) For any individual x, the expression "x is identical with x" is true,

is false. This is because "x is identical with x" is no more true or false than an out-of-the-blue utterance of "x is even." What *is* true is the statement that:

(6) A true sentence results if the occurrences of "x" in "x is identical with x" are replaced with an expression, the same one in both cases, of an expression that picks out some object.

The standard way of abbreviating (5) is:

(7) For any x, ‹x is identical with› is true.

Quasi-quotes will occur frequently in the chapters on the philosophy of language. It's not always easy to know when they occur; and I often lapse, using quotation marks instead of them, or just omitting them. When I do this, mentally insert them—otherwise what I'm saying will be false. (I appreciate it.) See "quotation marks."

Quotation marks: The word "snow" has four letters. But snow doesn't have four letters. So the sentence "snow has four letters" is false. Snow is white, crystalline substance, which consists, not of letter, but of H_2O molecules. In the philosophy of language, we often wish to refer to expressions themselves —we often wish to talk, not about snow (the substance), but "snow" (the word that substance).We do this by putting *quotation marks* around the expression to which we refer. So "snow" refers to the *word* that refers to the substance. See "quasi-quotation marks."

Rationalism: The doctrine that not all knowledge is derived from sense-perception, the reason being that some knowledge is acquired through the analysis of concepts. Rationalists typically hold that knowledge is a

prerequisite for our acquiring sensory information and also for our making any inferences from it. Since such knowledge couldn't possibly be observation based, it must have some other basis, this other basis being "pure reason." Rationalists hold that knowledge of logical norms is non-perceptual, their argument being that sense-perception apprises us of what is, not of what ought to be (even in a narrowly logical sense). See "empiricism."

Reductio ad absurdum: An argument that proves its point by showing that the negation of that point, if assumed true, leads to a contradiction (a statement of the form *Q and not-Q*). Suppose that P is true, that you know this, and wish to prove it to a friend of yours who doubts it. One way of doing so is to assume the negation of P and then show that has a self-contradictory result. Your position is that there are no even primes greater than two. The negation of your position is that there are even primes greater than two. You apprise your friend of your viewpoint. He disagrees. You respond by saying: "Let's suppose you're right. In that case, there is some number n that is greater than two and is even and is prime. Being prime, n is divisible only by itself and one. Being greater than two, $n \neq 2$. And being even, n is divisible by two. Thus, n is divisible by itself and one *and* by two. Therefore it isn't divisible *only* by itself and one. So n both is, and is not, divisible *only* by itself and one. You've reduced your friend's position to absurdity.

Refer: The word "refer" is ambiguous. It may refer to a property of actions or to a property of expressions. (In the present work, it is almost always used to denote a property of expressions.)

In an utterance of the sentence "when I was discussing the difference between real virtue and feigned virtue, I was referring to your father," the occurrence of the word "referring" picks out a characteristic of a certain ***speech act.*** Given some object O, a given act A that is performed by so and so is a case of so and so's referring to O if, in performing A, it was so and so's intention to make a statement about O. Let's refer to this sort of reference as "reference*."

In the sentence, "the name 'Socrates' refers to a certain philosopher," the occurrence of the word "refers" picks out a property of expressions, not of actions. To say that an expression E refers to some object O is to say that, in virtue of containing E, a sentence *ipso facto* concerns O. "Socrates" refers to

Socrates because, in virtue of having the form "...Socrates...," a sentence says that Socrates has...x...("Socrates is tall" says that Socrates has the property of being a thing x such that x is tall. "Plato admires Socrates" says that Socrates has the property of being a thing x such that Plato admires x.) Let us refer to this sort of reference as "reference#."

Some authors[445] have tried to show that reference# was identical with reference* or, at the very least, capable of being understood in terms of it. This is radically absurd.[446] Reference* presupposes reference#. Nobody could refer* to anything unless there were already meaningful expressions and, therefore, referring# terms. In Chapter 4, Sections 5.3.2–5.3.4, it is explained why this is so. (So far as I know, it was John Searle, in his classic (1966) work *Speech Acts*, who made this important point for the first time in print.)

Referent: the reference of an expression is what it refers to. See "refer."

Scope: See "modal operator."

Second-order desire: A desire about a desire (e.g., a desire to refrain from acting on one's desire to have a cigarette).

Self-contradictoriness: Any statement that has the form ‹P and not-P› or that implies a statement of that form is self-contradictory. All self-contradictory statements are false. Anything that entails a self-contradictory statement is false.

Semantic externalism: The doctrine that what our words and sentences mean isn't determined wholly by what our thoughts are, but also by environmental factors of which we may be unaware.

Semantic rule: A rule that assigns meaning to an expression. It's a semantic rule of English that "snow" refers to a certain crystalline white substance, and "water" refers to H_2O, etc.

At least some semantic rules assign meaning to expression-types, on the basis of which tokens of those types have this or that meaning depending on the circumstances. Thus, the meaning of the word-type "I" doesn't assign this or that individual to it. "I" per se doesn't refer to you or me or anyone else.

But “I” does have a meaning. That meaning is a rule to the effect that tokens of it refer to the people who produced those tokens. See “expression-type.”

Sentential force: Some sentences are true or false (e.g., “snow is white”). But some are neither true nor false. Imperatives (i.e., orders—e.g., “get out of here”) and interrogatives (i.e., questions—e.g., “what’s a hypotenuse?) are neither true nor false. True or false sentences are known as *indicatives*, and they therefore differ in respect of *force* from both interrogatives and imperatives.

Sense vs. reference: “The inventor of bifocals” refers to Benjamin Franklin, and so does “the first postmaster general.” But, even though those expressions co-refer, they have different *senses.* Although Benjamin Franklin was the first postmaster general, it isn’t in *virtue* of his having this property that “the inventor of bifocals” refers to him. And although Benjamin Franklin was the inventor of bifocals, it isn’t in *virtue* of his having this property that “the first postmaster general” refers to him. It is in virtue of his being a unique first postmaster general and a unique inventor of bifocals that Benjamin Franklin is the referent of, respectively, “the first postmaster general” and “the inventor of bifocals.” The *sense* of an expression is the property that an object must have to be its referent. So supposing that O is the referent of some expression R, and that it is in virtue of O’s having property P that R refers to O, P is the “sense” of R.

Commentary: Those who hold that definite descriptions are quantifiers, as opposed to referring terms, will reject what was just said. See Chapter 6. See “definite descriptions” and “quantifier.”

Sense-perception: Sight, hearing, touch, taste, smell, and one’s bodily kinesthetic sense (the feeling of resistance one experiences when trying to move objects—this sensory modality involves touch but isn’t identical with it). The two distinguishing features of sense-perception are (i) that it yields knowledge that is (at the ***personal level of cognition***) non-inferential, and (ii) perceptual information varies in real-time with the states of affairs it concerns. See “personal stratum of cognition.”

Sets and classes: A class is simply a collection of some kind—for example, the class of numbers, the class of even numbers, the class of people, the class

of people who are taking philosophy 101 this semester, the class of smart people, the class of talented athletes.

The word "set" is sometimes used instead of "class." We'll use these terms interchangeably. (Some set-theorists, e.g., Quine (1956), do not do so.)

Commentary: The sets that are of intellectual interest are usually ones whose members have some property in common with one another that nothing else has in common with any of them. So, for example, we talk about the set of whole numbers, the set of people who have survived pancreatic cancer, the set of people who are currently philosophy professors. Let S_1, S_2, and S_3 be these three sets. Every member of S_1 has the property of being a whole number and nothing that isn't a member of S1 has that property. Thus, the property of being a whole number *generates* S1. The property of being a survivor of pancreatic cancer and the property of being a philosophy professor generate S_2 and S_3, respectively.

Technically, given *any* set, there is *some* property that all and only its members have in common. But oftentimes these properties are very contrived. Let S be the set consisting of Smith and the number two. Everything in S has the property of being identical either with Smith or with the number, and nothing outside of that set has that property. But, of course, we're usually interested in sets that are generated by explanatorily more robust properties (e.g., the property of being a number).

Additional commentary: Sets ≠ properties: Some have held that sets are with properties. Their argument: "Smith is a person" (i.e., "Smith has the property of being a person") is equivalent with "Smith is a member of the class of people," and the same thing *mutatis mutandis* holds of any statement that attributes any property to anything.

But sets are not properties, and it's easy to identify the fallacy in the foregoing argument. "Smith is a person" is *not* equivalent "Smith is a member of the class of people." Let S be the class of people currently alive, and suppose that both Smith and Jones are members of S. Remember that sets are individuated by their memberships. Thus, S couldn't possibly exist unless Jones did as well. But Smith could exist and be a person in a world where Brown and, therefore, S failed to exist. So "Smith is a person" is not

equivalent with “Smith is a member of S.”

Sets are modally and temporally frozen. Properties are modally and temporally elastic. A given property may have different instances at different times and in different universes. (Some properties are modally frozen—e.g., the property of being an even number. But such properties are not representative of properties in general, and we may disregard them in what follows.) Thus, for any property phi (with the parenthetically noted exceptions) and any set S, ‹x has phi› is not equivalent with ‹x is a member of S.›

Simpliciter: This means *without qualification.* “There is no such thing as motion *simpliciter*—one must be moving *relative to this or that framework.*”

Singular term: Any terms refers to an individual (e.g. “Socrates,” “Shamu”). See “referring term.”

Slingshot-style arguments: Consider the following sentences:

(1) Mozart was a composer.

(2) The class of all things x such that ***Mozart was a composer and x is self-identical*** is identical with the class of all things x such that ***x is self-identical***.

(3) The class of all things x such that ***Bill Clinton was U.S. President in 1997 and x is self-identical*** is identical with the class of all things x such that ***x is self-identical***.

(4) Bill Clinton was U.S. President in 1997.

Here is the position of those who believe in the cogency of slingshot-style arguments:

> (SS[447]) (1) is logically equivalent with (2). Therefore, (1) and (2) must co-refer, supposing that either refers to anything. (It is being assumed, very reasonably, that logically equivalent expressions cannot have different referents.)
>
> (3) is what results when a referring term in (2) is replaced with a co-referring term. (“The class of all things x such that Mozart was a composer and x is self-identical” refers to the class of all self-identical

objects, and so too does "the class of all things x such that Bill Clinton was U.S. President in 1997 and x is self-identical.") Since intersubstituting co-referring terms preserves reference, it follows that (2) and (3) co-refer, supposing that either refers to anything.

(3) is equivalent with (4). Therefore, (3) and (4) must co-refer, supposing that either refers to anything.

Therefore, (1) and (4) must co-refer, supposing that they refer to anything.

If the occurrences in (1)–(4) "Mozart was a composer" and "Bill Clinton was U.S. President in 1997" are replaced with occurrences of *any* two true sentences, the argument that results is cogent. Thus, any two true sentences co-refer, supposing that any true sentence refers to anything. (For the same reason *mutatis mutandis*, any two false sentences co-refer, supposing that any false sentence refers to anything.)

The property of being true is the one semantically significant thing that (1) and (4) have in common. So, supposing that sentences refer to anything, (1) and (4) must refer to the property of being true, and true sentences must in general be assumed to refer to the property of being true.[448] (For the same reason *mutatis mutandis*, false sentences must refer to the property of being false.[449])

In general, a slingshot-style argument is one to the effect that, since logically equivalent expressions can't have different referents and since intersubstituting co-referring terms preserves truth-value, some class that would seem to comprise several objects (for example, the class of things that are referred by true sentences) in fact only comprises one object (for example, the property of being true).

SS is a poor argument that embodies a spurious understanding of reference, the same being true of every other slingshot-style argument. Suppose for argument's sake that "the class of all things x such that Mozart was a composer x is self-identical" is functioning as a device of reference. In that case, it's picking out the class of self-identical objects, and what (2) is saying about it is that it's identical with itself; in other words, (2), thus interpreted, is saying that:

(2*) the class of self-identical objects is identical with the class of self-identical objects.

But (2*) is not equivalent with (1). Therefore, if the definite descriptions in (1) and (2) are assumed to be referring terms—which is precisely what they *are* being assumed to be—(2) isn't equivalent with (1). For the same reason *mutatis mutandis*, (4) is not equivalent with (3) if "the class of all things x such that Bill Clinton was U.S. President in 1997 and x is self-identical" is taken to be a device of reference.

In order for (2) to be equivalent with (1), the former must be taken to say that:

(2#) there exists a class to which a thing belongs iff Mozart is a composer and that thing is self-identical; and there exists a class to which a thing belongs iff that thing is self-identical; and the first class is identical with the second class.

But if (2) is interpreted in this way, the definite description occurring in it is a quantifier, not a referring term. In which case, (3) is *not* what result when a referring term is replaced with a co-referring term. (It's what results when a non-referring term is replaced with a non-referring term.)

So SS is a failure.

Also, it's obvious that "snow is white" doesn't co-refer with "the property of being true" (or with "the *True*"). In virtue of having the form, "...the property of being true..." a sentence is about the property of being true. ("Everything Bill says has the property of being true" says of the property of being true that it is instantiated by everything that Bill says.) But in virtue of having the form "...snow is white..." a sentence is about the *proposition* that snow is white. ("If snow is white, then snow isn't green" says that the proposition that snow is white, if true, has the proposition that snow isn't green as a consequence. Of course, "if snow is white, then snow isn't green" can, like any other sentence, be interpreted as concerning the property of being true—it can be interpreted as saying that the property of being true is a characteristic of the proposition that snow isn't green if it's white. But it isn't *in virtue of* it's containing an occurrence of "snow is white" that "if snow is white, then snow isn't green" concerns that property.) So unless the concept

of reference is warped beyond all recognition—unless the term "reference" is redefined to as to validate this argument—what SS establishes, supposing its premises true, is false. In any case, we've seen independently of this that not all of it's premises are true.

But Donald Davidson took it for granted that SS is cogent.[450] And he argued that, given the (alleged) fact that SS is cogent, it follows that there is only one fact. His argument is simple. Put the expression "the fact that" in front of each of (1)–(4). The same principles that show that (1) and (4) must co-refer also shows that "the fact that Mozart was a composer" co-refers with "the fact that Bill Clinton was U.S. President in 1997)."

Davidson concluded, absurdly, that these two facts *are* identical—and that, in general, there is only one fact. Having, in his mind, established that there is only one fact, he argued that mental states are not representations. Since there is only one fact, Davidson argued, no two mental states represent different facts. On this basis, Davidson concluded that differences between mental states—between, for example, my perception of the piano in my living room and your perception of the ink-marks in front of you—are not representational differences. Why is this? Because, says Davidson, representations, supposing them to exist, are representations of facts; and, as we've just seen, there is only one fact. Therefore, Davidson concludes, my perception of the piano in my living room is *representationally* just like your perception of these ink-marks. Thus, the differences between those perceptions have nothing to do with any representational differences that there might be between them. Which means that, if there are any such differences, they are innocuous. Which, in its turn, means that *either* (i) that neither perception represents anything or (ii) that, if either perception does represent something, it is irrelevant to how that perception differs from other perceptions (or mental states of other kinds) and, therefore, to how what perception is. Which, give or take some unimportant nuances, means that neither represents anything.

Davidson's position (that no mental entity represents anything) is obviously false, and his argument for it fails for the same reason that SS fails. But Davidson's argument involves a blunder of which SS is innocent. For a mental state to be representational is not for there to be some fact that it represents. It is for that mental state to be to the effect that there exists some

fact. Suppose that I hallucinate that there is a pink elephant in my living room. (Assume that, in actuality, there is no such elephant.) My hallucination is representational. It isn't like a rock. But there is no elephant x such that my hallucination represents some pink elephant x such that x is in my living room. My hallucination is representational in the sense that it is to the effect that that there exists some pink elephant x such that x is in my living room. (See Chapter 1, Section 3.2 for further clarification.)

In his (ironically named) book *Facing the Facts*[451], Stephen Neale puts forth what he believes to be an improvement on SS.[452] Neale's argument involves the following four assumptions:

G_1: The sentence "Fa" stands for the same fact as the sentence "a=the unique thing x such that (x = a and Fx)." (In this context, "=" is to be taken to mean "is identical with.")

G_2: Any sentence can be put into subject-predicate form.

G_3: What a complex expression refers to is a function of what it's parts refer to.

G_4: Definite descriptions are referring terms. "The even prime" is an expression that picks out the number two.

Assume that for some individuals a and b, and some properties G and F, the following are true: "Fa," "Gb," and "a ≠ b"; and suppose that these sentences stand for facts F_1, F_2, and F_3, respectively.

Consider the following seven sentences.

1. Fa. By assumption.
2. a ≠ b. By assumption.
3. Gb. By assumption.
4. a = the unique thing x such that (x = a and Fx). Follows from 1, by G_1.
5. a = the unique thing x such that (x = a and x ≠ b). Follows from 2, by G_2.
6. b = the unique thing x such that (x = b and Gx). Follows from 3, by G_1.

7. b = the unique thing x such that (x = b and x ≠ a). Follows from 2, by G_1.

1 and 4 stand for the same fact, namely F1. 2, 5, and 7 all stand for the same fact, namely F_2. And 3 and 6 stand for the same fact, namely F_3. "The unique thing x such that (x = a and Fx)" co-refers with "the unique thing x such that (x = a and x ≠ b)", since both expression refer to a. Therefore, 5 is what results when the definite description in 4 is replaced with a co-referring definite description. Therefore, 4 and 5 refer to the same fact. Since they refer, respectively, to F_1 and F_2, it follows that F_1 and F_2 are the same fact. "The unique thing x such that (x = b and x ≠ a)" co-refers with "the unique thing x such that (x = b and Gx)," since both expressions refer to b. Therefore, 7 is what results when the definite description in 6 is replaced with a co-referring definite description. Therefore, 6 and 7 refer to the same fact. Since they refer, respectively, to F_3 and F_1, it follows that F_3 and F_1 are the same fact. Now, if G_4 is wrong, then definite descriptions are not referring terms, which case 5 is not what results when the definite descriptions in 4 is replaced with a co-referring expression, and 7 is what not results when the definite description in 6 is replaced with a co-referring expression. But assuming that G4 is right, we've established that $F_1 = F_2$. Therefore $F_1 = F_2 = F_3$, and "Fa" therefore co-refers with "Gb" and also with "a ≠ b." So, assuming that definite descriptions are what they appear to be, viz. referring terms, it follows there is only one fact.

As Neale points out, this result, if it is in fact true, is astonishing. But, as Neale then points out, G_2 allows us to go further. Let S and S* be *any* two distinct statements that are true. (So S and S* needn't have subject-predicate form.) According to G_2, S and be put in subject-predicate form, and so can S*. Given any existing individual a, S can be rendered as "a is an x such that S"; and given any existing individual b, S* can be rendered as "b is an x such that S*." And once that is granted, it follows, by the foregoing argument, that S and S* stand for the same fact. It follows that, given G_1-G_4, any two true sentences stand for the same fact.

Evaluating Neale's argument: The problems with this argument are very similar to the problems with Davidson's. Consider line 4 of Neale's

argument:

4. a = the unique thing x such that (x = a and Fx).

If this sentence is read ***attributively*** (in other words, if the definite descriptions in them are treated as quantifiers, as opposed to referring terms), then it is logically equivalent with

1. Fa,

in which case those two sentences can therefore be seen as standing for the same fact. But if 4 is read ***referentially*** (in other words, if the definite descriptions in them are treated as referring terms, as opposed to quantifiers), then 4 stands for the same fact as:

4_R: a = a,

in which case 1 and 4 cannot plausibly be seen as standing for the same fact. For exactly similar reasons, 5, 6, and 7 must be read attributively if they can plausibly be regarded as standing for the same facts as, respectively, 2, 3, and 4. In general, G_1 is false if 4–7 are read referentially, 4–7 must therefore be read attributively. Bottom line: Neale's argument fails if 4–7 are read referentially. It also fails, for obvious reasons, if they're read attributively.

Sorites paradox: Any paradox like the following. Somebody with zero dollars is poor. Giving one dollar to a poor person won't make him stop being poor. So, no matter how many times you give a dollar to a poor person, he won't stop being poor. Therefore, a person who has been given $500,000,000 dollars is poor. A sorites paradox always consists of a base-clause (e.g., "a person with zero dollars is poor"), an inductive clause (e.g., "if a person with n dollars is poor, so is a person with n + 1 dollars), and an obviously false conclusion that seems to follow (e.g., "a person with a billion dollars is poor")."

Commentary: My solution to this paradox: Statements of the form "x is rich" are elliptical for statements of the form "x is rich compared to y," where y is either some person or some benchmark. When the benchmark is identified, the underlined sentence is seen to be false. If Frank is the person/benchmark in question, and Frank has $1,000,000, then there is a number n (namely,

999,999.7) such that a person with n dollars is not rich and such a person with n dollars is rich.

Special compositionality: A special form of compositionality, which we will refer to as "special compositionality," concerns reference. According to special compositionality, what a referring term refers to is a function of what its parts refer to. "The author of *War and Peace*" refers to one person. "The author of *Crime and Punishment*" refers to someone else. Why do those expressions refer to different people? Because *War and Peace* doesn't refer to the same book as *Crime and Punishment.* This shows that what "The author of *War and Peace*" refers to depends on, and is thus "a function of," *War and Peace* refers to, and it shows the same thing *mutatis mutandis* to be true of "The author of *Crime and Punishment.*" Given obvious extensions of this argument, it follows that *in general* what referring terms refer to depends on what their parts refer to. See "compositionality."

Speech act[453]: Assertions are speech acts, and so are promises, orders, exhortations, and so on. A rough definition of "speech act" is "any case of somebody's deliberately uttering or otherwise ***tokening*** a sentence."

A person may produce a burst of noise, and *know* it to be a sentence, but not thereby have performed a speech act. Knowing that you find loud "s"-sounds completely intolerable, I shout out "snow sometimes symbolizes sorrow" while in your presence. Even though I knew the noises I was making to be a sentence-token, it wasn't my intention in producing them to token a sentence, and I therefore wasn't performing a speech act.

Surd: A "surd" is anything that cannot coherently be supposed to exist. Thus, square circles are surds, and so are events that pre-date themselves.

Though pedagogically effective, this definition is logically defective. There exists nothing that cannot possibly exist. (There exists no object x such that x cannot possibly exist. Thus, there is nothing to which, in saying that "surds" are things that cannot possibly exist, we are saying that "surd" refers. (There exists no x such that, in saying that the word "surd" refers to x, we are saying that "surd" refers.)

The word "surd" can be non-defectively defined only in a circuitous manner. For any property phi, the sentence ‹anything having phi is a surd,›

means that the laws of logic prohibit phi from being instantiated. (Given that not all sentences containing the word "surd" have the form ‹anything having phi is a surd,› this definition needs fine-tuning. But whereas the problems with our initial definition were substantive, the problems with this definition are merely technical.)

Supervenience: Two organisms that are microphysically exactly alike are also biologically exactly alike. Thus, the microphysical locks the biological into place. Put another way, the biological supervenes on the microphysical.

Two statues that are physically exactly alike are *ipso facto* aesthetically exactly alike. Thus, the aesthetic *supervenes* on the physical. Two situations that are physically and psychologically are *ipso facol* morally exactly alike. Thus, the moral supervenes on the physical and the psychological (taken together).

Synthetic knowledge: Knowledge is synthetic when what one knows is given by a synthetic truth. See "synthetic truth."

Synthetic truth: A statement is synthetic if it isn't analytic. See "analytic" and "synthetic knowledge."

Tautology/truism: A tautology (or truism) is a trivially true statement—for example, "there are three feet in a yard" or "Tuesdays come after Mondays" or "there are two weeks in a fortnight."

Theory of Descriptions: The position, first advocated in 1905 by Bertrand Russell, that definite descriptions are not devices of reference and are instead quantifiers. "The inventor of bifocals" seems to refer to Benjamin Franklin. But "the inventor of bifocals snored" is true iff exactly one thing invented bifocals, and any such snored. "Smith kicked the inventor of bifocals is true iff exactly one thing invented bifocals, and Smith kicked any such thing. What this shows is that, in virtue of containing an occurrence of "the inventor of bifocals," an expression encodes an existence-claim. This, in its turn, suggests that "the inventor of bifocals," and by parity of reasoning all other definite descriptions, are quantifiers, not singular terms.

Truth: A true proposition is said to be one that "corresponds to the facts." But this statement is useless, since the expression "corresponds to the facts"

is as obscure as the expression “true.”

Given any proposition, there is some condition that is fulfilled if, and only if, that proposition is true. Conditions may be identified with properties. Therefore, truth may be identified with instantiatedness, and propositions (i.e., the things that have the property of being true) are properties (or, strictly speaking, sets there). See Chapter 3.

Many philosophers (e.g., Ayer, Hempel, Strawson, Wittgenstein) were, at least at certain junctures in their careers, of the view that “truth” (and “true” and other such terms) were meaningless. There is, it was said, no difference between saying “the sentence ‘snow is white’ is true” and “snow is white”; therefore, the words “it is true that” (and “it is a truth that,” and so on) are useless. Wittgenstein was vehement in his advocacy of this viewpoint, which is known as “deflationary” (or the “redundancy”) analysis of truth. The deflationary analysis is false. If “white” meant “black,” but the English language were otherwise unchanged, “snow is white” would be true in a world where snow was black and false in a world where snow was white. So

(1) “snow is white”

is not equivalent with

(2) “the sentence ‘snow is white’ is true.”

(1) makes a statement about snow; it doesn’t say anything about language. (2) makes a statement about language; it doesn’t say anything about snow. (2) concerns the *word* “snow.” But snow isn’t a word; it’s a substance.

This brings us to next problem with the deflationary analysis. (2) is a true statement. There is no way to make that statement without using the word “true” or some equivalent. If you just say “snow is white,” you are not, as we just say, saying anything about the word “snow.” (What you are saying, in saying “snow is white,” would be true in a world where there were no words, but where snow was white.)

Also, any sentence containing a quantifier contains an occurrence, if only a phonetically unrealized one, of the word “true” or of some equivalent. To say

(SD) “some dolphins can read”

is to say that

(SD) for some object x, the proposition *x can read* is true.

Truth-value: A true sentence has the truth-value *true*, and a false sentence has the truth-value *false.* The truth-values *true* and *false* can be thought of as, respectively, the property of being true and the property of being false.

If two sentences are both true or they are both false, they have the same "truth-value." So "1 + 1 = 2" and "2 + 2 = 4" have the same truth-value—they have truth-value *true*. And so "1 + 1 = 3" and
"2 + 2 = 5" have the same truth-value—the truth-value *false.*

Universal generalization: A sentence expresses a "universal generalization" if the ***main connective*** of that sentence is a ***universal quantifier.***[454] A connective is an expression that, when given a sentence, or open sentence, or ordered n-tuple of (open) sentences, yields a new sentence. Thus "and" is a connective because, when given the pair of sentences ("grass is green," "snow is white"), it yields the sentence "grass is green and snow is white." (In what follows, "sentence" will be short for "sentence or open sentence of ordered n-tuple of sentences or open-sentences.")

Commentary: Many statements that *seem* to be conditionals are not, and are instead universal generalizations. For example *if a thing likes apples, it is friendly*, is a universal generalization, not a conditional. The reason: it's identical with the proposition: *all things are such that they are friendly if they like apples.* Put another way, it's identical with: *for any x, if x likes apples, x is friendly.* Thus, the "main connective" of the sentence "if a thing likes apples, it is friendly" isn't "if," and is instead "for any object x"; and for this reason, it's a universal generalization, not a conditional.

Utilitarianism: The doctrine that an act is morally correct to the extent that it promotes human welfare, and morally incorrect to the extent that it undermines it. See "consequentialism."

Vagueness: See "ambiguity vs. vagueness."

Valid: An argument is valid iff the conclusion cannot possibly be false, if premises are true. A valid argument can have false premises. (For example: Premise: All people have feathers. Premise: All things that have feathers drink to excess. Conclusion: All people drink to excess.) A valid argument

with true premises is a *sound* argument. (For example: Premise: JMK has a car. Premise: All cars are motorized vehicles. Conclusion: JMK has a motorized vehicle.)

Validate: See "interpretation."

Virtue theory: The doctrine composed of the following two assertions: (i) a morally correct life is one in which one flourishes. (ii) One's own flourishing depends constitutively on one's having character traits (or "virtues") that conduce to the flourishing of others.

BIBLIOGRAPHY

Armstrong, David. 1989. *Universals: an Opinionated Introduction.* Boulder, CO: Westview Press.

Armstrong, David. 1992. Properties. In Steven M. Cahn (Ed.), *Philosophy for the 21st Century* (pp. 181–193). New York: Oxford University Press.

Austin, J. L. 1961. *Philosophical Papers.* New York: Oxford University Press.

Austin, J. L. 1976. *How to do things with words.* New York: Oxford University Press.

Ayer, Alfred Jules. 1952. *Language, Truth, and Logic.* New York: Dover Publications.

Ayer, Alfred Jules. 1954. Can there be a private language? In A.P. Martinich (Ed.), *The Philosophy of Language* (pp. 449–456). Oxford: Oxford University Press.

Barwise, Jon. 1989. *The Situation in Logic*. Palo Alto, CA: CSLI Publications.

Barwise, Jon, & Perry, John. 1983. *Situations and Attitudes.* Palo Alto, CA: CSLI Publications.

Benacerraf, Paul. 1965. What Numbers Could Not Be. *Philosophical Review,* 74, 47–73.

Berkeley, George. 1934. *An Essay Towards a New Theory of Vision.* Oxford: Clarendon Press.

Bezuidenhout, A., et al. (Eds.). *Descriptions and Beyond.* New York: Oxford University Press.

Blackburn, Simon. 1984. *Spreading the Word.* Oxford: Clarendon.

Bonjour, Laurence. 1985. *The Structure of Empirical Knowledge.* Cambridge, MA: Harvard University Press.

Bonjour, Laurence. 1998. *In Defense of Pure Reason.* Cambridge, U.K.: Cambridge University Press.

Brandom, Robert. 1994. *Making it Explicit.* Cambridge, MA: Harvard University Press.

Burge, Tyler. 1979. Individualism and the Mental. In A. Pessin and S. Goldberg (Eds.), *The Twin-Earth Chronicles,* (pp. 125–141). Armonk, NY: M.E. Sharpe.

Burge, Tyler. 1980. Computer proof and a priori knowledge. *Journal of Philosophy* 77, 797–803.

Burge, Tyler. 1982. Other Bodies. In A. Pessin and S. Goldberg (Eds.), *The Twin-Earth Chronicles,* (pp. 142–160). Armonk, NY: M.E. Sharpe.

Burge, Tyler. 1986. Individualism and Self-knowledge. In A. Pessin and S. Goldberg (Eds.), *The Twin-Earth Chronicles,* (pp. 342–354). Armonk, NY: M.E. Sharpe.

Cain, M. J. 2002. *Fodor: Language, Mind and Philosophy.* Cambridge, MA: Polity Press.

Cappelen, Herman, & Lepore, Ernie. 2005. *Insensitive Semantics: A defense of speech act pluralism and semantic minimalism.* Oxford: Blackwell.

Carnap, Rudolph. 1932. The Elimination of Metaphysics through Logical Analysis of Language. In A.J. Ayer (Ed.), *Logical Positivism,* (pp. 60–81). New York: The Free Press.

Carnap, Rudolph. 1933. Psychology in Physical Language. In A.J. Ayer (Ed.), *Logical Positivism,* (pp. 165–198). New York: The Free Press.

Carnap, Rudolph. 1934. *The Unity of Science.* London: K. Paul, Trench, Kubner & Co, Ltd.

Carnap, Rudolph. 1937. *The Logical Syntax of Language.* Routledge & Kegan Paul, London.

Carnap, Rudolph. 1956. *Introduction to Symbolic Logic.* New York: Dover.

Carnap, Rudolph. 1966. *Introduction to the Philosophy of Science.* New York: Dover.

Chalmers, David. 1996. *The Conscious Mind.* New York: Oxford University Press.

Chisholm, Roderick. Human Freedom and the Self. In S. Cahn (Ed.), *Philosophy for the 21st Century,* (pp.407–413). New York: Oxford

University Press.

Chomsky, Noam. 1980. *Rules and Representations.* New York: Columbia University Press.

Chomsky, Noam. 1988. Language and the Problems of Knowledge. In A.P. Martinich (Ed.), *The Philosophy of Language,* (pp. 509–527). New York: Oxford University Press.

Church, Alonzo. 1958. The Ontological Status of Women and Abstract Entities. Lecture delivered at Harvard, available online at: *http://www.cs.nyu.edu/pipermail/fom/2005-September/009079.html*

Churchland, Paul. 1984. *Matter and Consciousness.* Cambridge, MA: The MIT Press.

Cresswell, Max. 1985. Structured Meanings. In J. Garfield and M. Kiteley (Eds.), *Meaning and Truth,* (pp. 446–452). St. Paul, MN: Paragon Press.

Davies, Peter J. 2001a. *The Character of a Genius: Beethoven in Perspective.* Westport, CT: Greenwood Press.

Davies, Peter J. 2001b. *Beethoven in Person: His deafness, illness, and death.* Westport, CT: Greenwood Press.

Davidson, Donald. 1967. Truth and Meaning. *Synthese* 17, 304–323.

Davidson, Donald. 1980. *Inquiries into Truth and Interpretation.* New York: Oxford University Press. Davidson, Donald. Truth and Predication. Cambridge, MA: Harvard University Press.

Dennett, Daniel. 1975. Eliminative materialism and the propositional attitudes. In In D. Rosenthal (Ed.), *The Nature of Mind,* (pp. 502–507). New York: Oxford University Press.

Dennett, Daniel. 1978. *Brainstorms.* Cambridge, MA: The MIT press.

Donnellan, Keith. 1958. Reference and Definite Descriptions. In A.P. Martinich, (Ed.), *The Philosophy of Language* (pp. 235–247). New York: Oxford University Press.

Donnellan, Keith. 1966. Reference and Definite Descriptions. In A.P. Martinich (Ed.), *The Philosophy of Language,* (pp. 235–257). Oxford: Oxford University Press.

Donnellan, Keith. 1974. Speaking of nothing. *The Philosophical Review* 74, 3–31.

Dretske, Fred. 1982. *Knowledge and the Flow of Information.* Cambridge, MA: The MIT Press.

Ducasse, C.J. 1969. *Truth, Knowledge, and Causation.* London: Routledge & Kegan Paul.

Dummett, Michael. 1973. *Frege: Philosophy of Language.* Cambridge, MA: Harvard University Press.

Dummett, Michael. 1978. *Truth and Other Enigmas.* Cambridge, MA: Harvard University Press.

Einstein, Albert. 1962. *The Principles of Relativtiy.* New York: Dover.

Einstein, Albert, & Infeld, Leopold. 1961. *The Evolution of Physics.* New York: Simon & Schuster. Evans, Gareth. 1982. *The Varieties of Reference.* Oxford: Clarendon Press.

Evans, Gareth. 1985. *Collected Papers.* Oxford: Clarendon Press.

Falvey, Kevin. 1994. Externalism, Self-Knowledge, and Skepticism. Unpublished Dissertation. Department of Philosophy. University of Minnesota.

Falvey, Kevin, & Owens, Joseph. 1994. Externalism, Self-Knowledge, and Skepticism. *Philosophical Review* 103, 107–137.

Feldman, Fred. 1978. *Introductory Ethics.* Englewood Cliffs, NJ: Prentice Hall.

Field, Hartry. 1977. Logic, Meaning, and Conceptual Role. *Journal of Philosophy* 69, 378–408. Fodor, Jerry. 1968. *Psychological Explanation.* New York: Random House.

Fodor, Jerry. 1975. *The Language of Thought.* New York: Thomas Y. Crowell.

Fodor, Jerry. 1981a. Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology. In D. Rosenthal (Ed.), *The Nature of Mind,* (pp. 485–498). New York: Oxford University Press.

Fodor, Jerry. 1981b. *Representations.* Cambridge, MA:

The MIT Press.

Fodor, Jerry. 1983. *The Modularity of Mind.* Cambridge, MA: The MIT Press.

Fodor, Jerry. 1987. *Psychosemantics.* Cambridge, MA: The MIT Press.

Fodor, Jerry. 1987b. Individualism and Supervenience. In A. Pessin and S. Goldberg (Eds.), *The Twin-Earth Chronicles,* (pp. 192–218). Armonk, NY: M. E. Sharpe.

Fodor, Jerry. 1990. *A Theory of Content and Other Essays.* Cambridge, MA: The MIT Press.

Fodor, Jerry. 1994a. Fodor. In S. Guttenplan (Ed.), *A Companion to the Philosophy of Mind,* (pp. 292–300). Oxford: Blackwell.

Fodor, Jerry. 1994b. *The Elm and the Expert: Mentalese and Its Semantics.* Cambridge, MA: The MIT Press.

Fodor, Jerry. 1998a. *Concepts.* Oxford: Clarendon Press.

Fodor, Jerry. 1998b. *In Critical Condition.* Cambridge MA: The MIT Press.

Fodor, Jerry, & Lepore, Ernest. 2002. *The Compositionality Papers.* New York: Oxford University Press.

Fodor, Jerry, & Pylyshin, Zenon. 1988. Connectionism and cognitive architecture: a critical analysis. In S. Pinker (Ed.), *Connections and Symbols,* (pp. 3–72). Amsterdam: Elsevier.

Frank, Philipp. 1949. *Philosophy of Science: the link between philosophy and science.* New York: Dover.

Frankfurt, Harry. 1988. *The Importance of What We Care About.* Cambridge, U.K.: Cambridge University Press.

Frege, Gottlob. 1879. Concept-Writing. In Michael Beaney (Ed.), *The Frege Reader* (pp. 181–193). Oxford: Blackwell.

Frege, Gottlob. 1891. Function and Concept. In Michael Beaney (Ed.), *The Frege Reader,* (pp. 130–148). Oxford: Blackwell.

Frege, Gottlob. 1892a. On *Sinn* and *Bedeutung.* In Michael Beaney (Ed.), *The Frege Reader,* (pp. 151–171). Oxford: Blackwell.

Frege, Gottlob. 1892b. On Concept and Object. In Michael Beaney (Ed.), *The*

Frege Reader, (pp. 181–193). Oxford: Blackwell.

Frege, Gottlob. 1918. Thought. In Michael Beaney (Ed.), *The Frege Reader,* (pp. 325–345). Oxford: Blackwell.

Freud, Sigmund. 1915. The unconscious. In P. Rieff (Ed.), *General Psychological Theory,* (pp. 116–150). New York: Macmillan Publishing Co.

Freud, Sigmund. 1965. *New Introductory Lectures on Psychoanalysis.* London: W.W. Norton and Co. Freud, Sigmund. 1989a. *An Outline of Psychoanalysis.* New York: W.W. Norton and Co.

Freud, Sigmund. 1989b. *Civilization and its Discontents.* London: W.W. Norton and Co.

Freud, Sigmund. 1998. *The Interpretation of Dreams.* New York: Basic Books.

Fuller, Lon. 1964. *The Morality of Law.* New Haven, CT: Yale University Press.

Gardner, Martin. 1976. *The Relativity Explosion.* New York: Vintage Books.

Gettier, Edmund. 1963. Is knowledge justified true belief. *Analysis,* 23, 121–123.

Gödel, Kurt. 1953. Is mathematics syntax of language? In S. Feferman (Ed.), *Kurt Gödel: Collected Works, Volume III,* (pp. 334–356). New York: Oxford.

Goodman, Nelson. 1954. *Fact, Fiction, and Forecast.* Cambridge, MA: Harvard University Press.

Goodman, Nelson. 1976. *The Languages of Art.* Indianapolis, IN: Hackett.

Grice, H.P. 1957. Meaning. In A.P. Martinich (Ed.), *The Philosophy of Language,* (pp. 72–78). Oxford: Oxford University Press.

Hacker, P.M.S. 1999. *Wittgenstein on Human Nature.* New York: Routledge.

Hacker, P.M.S. 1996. *Wittgenstein: Mind and Will.* Oxford: Blackwell.

Hacker, P.M.S., & Baker, Gordon. 1980. *Wittgenstein: Understanding and Meaning.* Oxford: Blackwell.

Hacker, P.M.S., & Baker, Gordon. 1984a. *Scepticism, Rules, and Language.* Oxford: Blackwell.

Hacker, P.M.S., & Baker, Gordon. 1984b. *Language, Sense, and Nonsense.* Oxford: Blackwell.

Hacker, P.M.S., & Baker, Gordon. 1985. *Wittgenstein: Rules, Grammar, and Necessity.* Oxford: Blackwell.

Hacking, Ian. 1975. *Why does language matter to philosophy?* Cambridge, U.K.: Cambridge University Press.

Hahn, Hans. 1933. Logic, Mathematics, and Knowledge of Nature. In A.J. Ayer (Ed.), *Logical Positivsim,* (pp. 137–163). New York: The Free Press.

Hale, Bob, & Wright, Crispin (Eds.). 1997. *A Companion to the Philosophy of Language.* Oxford: Basil Black-well.

Hare, R.D. 1999. *Without Conscience.* New York: The Guilford Press.

Hare, R.M. 1963. *Freedom and Reason.* New York: Oxford.

Hart, H.L. A. 1961. *The Concept of Law.* New York: Oxford.

Hempel, Carl. G. 1945a. Studies in Logic and Confirmation. *Mind,* 54, 1–26.

Hempel, Carl. G. 1945b. Studies in Logic and Confirmation. II. *Mind,* 54, 97–121.

Hempel, Carl G. 1950. The Empiricist Criterion of Meaning. In A.J. Ayer (Ed.), *Logical Positivsim,* (pp. 108–129). New York: The Free Press.

Hempel, Carl G. 1952. *Fundamentals of Concept Formation in Empirical Science.* Chicago: University of Chicago Press.

Hempel, Carl G. 1965. *Aspects of Scientific Explanation.* New York: The Free Press.

Hempel, Carl G. 1966. *Philosophy of Natural Science.* Englewood

Cliffs: Prentice Hall.

Jackson, Frank. 1998. *Mind, Method, and Conditionals.* London: Routledge.

Jackson, Frank, & Pettit, Philip. 2004a. Causation in the Philosophy of Mind. In F. Jackson, P. Pettit, and M. Smith (Eds.), *Mind, Morality, and Explanation,* (pp. 45–68). Oxford: Oxford University Press.

Jackson, Frank, & Pettit, Philip. 2004b. Some content is narrow. In F. Jackson, P. Pettit, and M. Smith (Eds.), *Mind, Morality, and Explanation,* (pp. 69–94). Oxford: Oxford University Press.

Jackson, Frank, & Pettit, Philip. 2004c. Functionalism and Broad Content. In F. Jackson, P. Pettit, and M. Smith (Eds.), *Mind, Morality, and Explanation,* (pp. 95–118). Oxford: Oxford University Press.

Jackson, Frank, & Pettit, Philip. 2004d. Program Explanation: A General Perspective. In F. Jackson, P. Pettit, and M. Smith (Eds.), *Mind, Morality, and Explanation,* (pp. 119–130). Oxford: Oxford University Press.

Hershenov, David. 2002. *A Defense of the Biological Approach to Personal Identity.* Unpublished dissertation. Department of Philosophy, University of California at Santa Barbara.

Hershenov, David. 2005. Do Dead Bodies Pose a Problem for Biological Approaches to Personal Identity? *Mind* 114, 31–59.

Hilbert, David. 1999. *Foundations of Geometry.* Chicago: Open Court.

Horst, Stephen. 1996. *Symbols, Computation, and Intentionality.* Berkeley, CA: University of California Press.

Kaplan, David. 1968. Quantifying In. In A.P. Martinich (Ed.), *The Philosophy of Language* (pp. 362–391). New York: Oxford University Press.

Kaplan, David. 1989a. Demonstratives. In J. Almog, et al. (Eds.), *Themes from Kaplan* (pp. 481–563). New York: Oxford University Press.

Kaplan, David. 1989b. Afterthoughts. In J. Almog, et al. (Eds.), *Themes from Kaplan* (pp. 565–614). New York: Oxford University Press.

Katz, Jerrold. 1972. *Semantic Theory.* New York: Harper and Row.

Kernberg, Otto. 1985. *Aggression in Personality Disorders and Perversions.* New Haven, CT: Yale University Press.

Kernberg, Otto. 1993. *Severe Personality Disorders.* New Haven, CT: Yale University Press.

Kim, Jaegwon. 1993. *Supervenience and Mind.* Cambridge, U.K.: Cambridge University Press.

King, Jeffrey C. 1995. Structured Propositions and Complex Predicates, *Nous,* 29(4), 516–535.

King, Jeffrey C. 1996. Structured Propositions and Sentence Structure, *Journal of Philosophical Logic,* 25: 495–521.

Kripke, Saul. 1979. A Puzzle about Belief. In A. Margalit (Ed.), *Meaning and Use.* Dordrecht: D. Reidel.

Kripke, Saul. 1980. *Naming and Necessity.* Cambridge, MA: Harvard University Press.

Kripke, Saul. 1982. *Wittgenstein on Rules and Private Language.* Cambridge, MA: Harvard University Press.

Kuczynski, John-Michael. 2007. *Conceptual Atomism and the Computational Theory of Mind: a defense of content-internalism and semantic externalism.* Amsterdam: John Benjamins.

Kuczynski, John-Michael. 2006a, "Does Possible World Semantics Turn all Propositions into Necessary ones?" *Journal of Pragmatics,* 39, (5): 872–916.

Kuczynski, John-Michael. 2006b. Two Concepts of Sentential "Form" and the So-called Computational Theory of Mind. *Philosophical Psychology,* 19, (6): 1–27.

Kuczynski, John-Michael. 2006c. Formal Operations and Simulated Thought. *Philosophical Explorations,* 9(2), 221–234.

Kuczynski, John-Michael. 2006d. On Marga Reimer's *Descriptions and Beyond. Pragmatics and Cognition,* 14(1), 196–204.

Kuczynski, John-Michael. 2005. The Concept of a Symbol and the Vacuousness of the Symbolic Conception of Thought. *Semiotica,* 154(4),

243–264.

Kuczynski, John-Michael. 2005b. Must one Know a Language to Grasp Propositions? *Teorema,* 24(2), 43–65.

Kuzynski, John-Michael. 2005c. *Literal Meaning and Cognitive Content.* Unpublished manuscript.

Kuczynski, John-Michael. 2004a. A non-Russellian Treatment of the Referential-Attributive Distinction. *Pragmatics and Cognition.* 12(2), 253–294.

Kuczynski, John-Michael. 2004b. Why Definite Descriptions Really Are Referring Expressions. *Grazer Philosophische Studien,* 68, 45–79.

Kuczynski, John-Michael. 2004c. Another Argument against the Thesis that there is a Language of Thought. *Communication and Cognition,* 37 (2), 83–104.

Kuczynski, John-Michael. 2004d. Non-declarative sentences and the theory of descriptions. *Principia,* 8(1), 119–154.

Kuczynski, John-Michael. 2003. Some Arguments against Intentionalism. *Acta Analytica,* 19(32), 107–142.

Kuczynski, John-Michael. 2002. Does the Idea of a "Language of Thought" Make Sense? *Communication and Cognition,* 35(4), 173–192.

Langford, C. H. 1942. The notion of analysis in Moore's philosophy. In P. A. Schilpp (Ed.), *The Philosophy of G.E. Moore,* (pp. 321–342). LaSalle: Open Court.

Lewicki, Pawel, Hill, Thomas, & Czyzewska, Maria. 1992. Nonconscious acquisition of information. *American Psychologist,* 47(6), 796–801.

Lewis, Clarence I. 1946. *An Analysis of Knowledge and Valuation.* Cambridge, MA: Harvard University Press.

Lewis, Clarence I. 1952. The modes of meaning. In L. Linsky (Ed.), *Semantics and the Philosophy of Language,* (pp. 50–66). Chicago: University of Illinois Press.

Lewis, David. 1975. Language and Languages. In A.P. Martinich (Ed.), *The Philosophy of Language,* (pp. 489–508). New York: Oxford University

Press.

Lewis, David. 1983. *Philosophical Papers, Volume 1.* New York: Oxford University Press.

Lewis, David. 1984. *On the Plurality of Worlds.* Oxford: Blackwell.

Lewis, David. 1987. *Philosophical Papers, Volume 2.* New York: Oxford University Press.

Lowe, Chelsea. 2007. *The Everything Guide to ODC.* Chicago: Adams Press.

Lycan, William. 1984. *Logical Form in Natural Language.* Cambridge, MA: The MIT Press.

Lyons, John. 1977. *Semantics.* Cambridge, U.K.:
Cambridge University Press.

Mackie, John Leslie. 1980. *The Cement of the Universe.* Oxford: Clarendon.

Mates, Benson. 1952. Synonymy. In L. Linskey (Ed.), *Semantics and the Philosophy of Language,* (pp. 111–138). Chicago: University of Illinois Press.

McCawley, James D. 1971. *Everything that Linguists Have Always Wanted to Know about Logic.* Chicago: University of Chicago Press.

McDowell, John. 1994. *Mind and World.* Cambridge, MA: Harvard University Press.

McDowell, John. 1998. *Meaning, Knowledge & Reality.* New York: Oxford University Press.

McWilliams, Nancy. 1994a. *Psychoanalytic Diagnosis: Understanding Personality Structure in the Clinical Process.* New York: The Guilford Press.

McWilliams, Nancy. 1994b. *Psychoanalytic Psychotherapy: A Practitioner's Guide.* New York: The Guilford Press.

Merleau-Ponty, Maurice. 1962. *The Phenomenology of Perception.* In Robert C. Solomon (Ed.), *Phenomenology and Existentialism,* (pp. 460–465). Boulder, CO: Rowman and Littlefield.

Merricks, Trenton. 2001. *Objects and Persons.* New York: Oxford University Press.

Montague, Richard. 1974a. The Proper Treatment of Quantification in Ordinary English. In Richard Thomason (Ed.), *Formal Philosophy.* New Haven, CT: Yale University Press.

Montague, Richard. 1974b. *Formal Philosophy: Selected Papers of Richard Montague.* New Haven, CT: Yale University Press.

Moore, G.E. 1912. Ethics. Available online at: http://fair-use.org/g-e-moore/ethics/

Moore, Robert C. 1995. *Logic and Representation.* Palo Alto, CA: CSLI Publications.

Nagel, Ernest. 1962. *The Structure of Science.* Indianapolis, IN: Hackett.

Nagel, Ernest. 2001. *Gödel's Proof.* New York: New York University Press.

Neurath, Otto. 1932. Protocol Sentences. In A.J. Ayer (Ed.), *Logical Positivism* (pp. 199–208). New York: The Free Press.

Nietzsche, F. 1966. *Beyond Good and Evil.* New York: Vintage.

Nietzsche, F. 1967a. *Genealogy of Morals.* New York: Vintage.

Nietzsche, F. 1967b. *The Will to Power.* New York: Vintage.

Pap, Arthur. 1949. *Elements of Analytic Philosophy.* New York: Macmillan.

Pap, Arthur. 1958. *Semantics and Necessary Truth.* New Haven, CT: Yale University Press.

Pap, Arthur. 1962. *Introduction to the Philosophy of Science.* New Haven, CT: Yale University Press.

Peirce, Charles S. 1955. *Selected Writings.* New York: Dover.

Plantinga, Alvin. 1974. *The Nature of Necessity.* New York: Oxford University Press.

Popper, Karl. 1957. *The Poverty of Historicism.* London: Taylor

& Francis.

Popper, Karl. 1959. *The Logic of Scientific Discovery.* London: Taylor & Francis.

Popper, Karl. 1983. *Realism and the Aim of Science.* London: Taylor & Francis.

Popper, Karl. 1992. *The Open Universe.* London: Taylor & Francis.

Putnam, Hilary. 1975. The Meaning of "Meaning." In A. Pessin and S. Goldberg (Eds.), *The Twin-Earth Chronicles,* (pp. 1–52). Armonk, NY: M.E. Sharpe.

Putnam, Hilary. 1996. Introduction to *The Twin-Earth Chronicles.* In A. Pessin and S. Goldberg (Eds.), *The Twin-Earth Chronicles,* (pp. xv-xxii). Armonk, NY: M.E. Sharpe.

Pylyshin, Zenon W. 1984. *Computation and Cognition.* Cambridge, MA: The MIT Press.

Quine, Willard van Orman. 1940. *Mathematical Logic.* Cambridge, MA: Harvard University Press, pp. 139–151.

Quine, Willard Van Orman. 1951. Two dogmas of empiricism. In A.P. Martinich (Ed.), *The Philosophy of Language,* (pp 26–39). Oxford: Oxford University Press.

Quine, Willard van Orman. 1953a. Reference and Modality. In W.V.O. Quine (Ed.), *From a Logical Point of View* (pp. 139–151). Cambridge, MA: Harvard University Press.

Quine, Willard van Orman. 1953b. *From a Logical Point of View.* Cambridge, MA: Harvard University Press.

Quine, Willard van Orman. 1956. *Set Theory and Its Logic.* Cambridge, MA: Harvard University Press

Quine, Willard van Orman. 1960. *Word and Object.* Cambridge, MA: The MIT Press.

Quine, Willard van Orman. 1966. *The Ways of Paradox and Other Essays.* Cambridge, MA: Harvard University Press.

Quine, Willard van Orman. 1970. *Philo sophy of Logic.* Cambridge, MA: Harvard University Press.

Quine, Willard van Orman. 1977. *Ontological relativity and other essays.* New York: Columbia University Press.

Quine, Willard van Orman. 1981. *Theories and Things.* Cambridge, MA: Harvard University Press.

Rachels, James. 1986. *The Elements of Moral Philosophy.* New York: Random House.

Ramsey, Frank. 1990. *Philosophical Papers.* Cambridge, U.K.: Cambridge University Press.

Reichenbach, Hans. 1947. *Elements of Symbolic Logic.* New York: Dover.

Reichenbach, Hans. 1957. *The Philosophy of Space and Time.* New York: Dover.

Richard, Mark. 1991. *Propositional Attitudes.* Cambridge, U.K.: Cambridge University Press.

Rorty, Richard. 1979. *Philosophy and the Mirror of Nature.* Princeton:

Princeton University Press.

Russell, Bertrand. 1903. *Principles of Mathematics.* Cambridge, U.K.: Cambridge University Press.

Russell, Bertrand. 1905. On Denoting. In A.P. Martinich (Ed.), *The Philosophy of Language,* (pp. 203–211). *The Philosophy of Language.* Oxford: Oxford University Press.

Russell, Bertrand. 1908. Mathematical Logic as Based on the Theory of Types. *American Journal of Mathematics, 30,* 222.

Russell, Bertrand. 1912. *The Problems of Philosophy.* London: George Allen Unwin.

Russell, Bertrand. 1917. *Mysticism and Logic.* London: George Allen Unwin.

Bertrand, Russell. 1918. The philosophy of logical atomism. In R.C. Marsh (ed.), *Logic and Knowledge,* (pp. 175–282).

Russell, Bertrand. 1920. *Introduction to Mathematical Philosophy.* London: George Allen Unwin.

Russell, Bertrand. 1921. *The Analysis of Mind.* London: George Allen Unwin.

Russell, Bertrand. 1927. *An Outline of Philosophy.* London: George Allen Unwin.

Russell, Bertrand. 1928. *The Analysis of Matter.* London: George Allen Unwin.

Russell, Bertrand. 1948. *Human Knowledge: Its Scope and Limits.* London: George Allen Unwin.

Russell, Bertrand. 1950. Logical positivism. In R.C. Marsh (Ed.), *Logic and Knowledge,* (pp. 367–382).

Russell, Bertrand. 1959. *My Philosophical Development.* London: George Allen Unwin.

Russell, Bertrand. 1984. *Theory of Knowledge.* London: George Allen Unwin.

Salmon, Nathan. 1986. *Frege's Puzzle.* Cambridge, MA:

The MIT Press.

Salmon, Nathan. 2005. *Metaphysics, Mathematics, and Meaning. Philosophical Papers II.* New York: Oxford University Press.

Salmon, Nathan. 2007. *Content, Cognition, and Communication: Philosophical Papers II.* New York: Oxford University Press.

Salmon, Wesley. 1984. *Scientific Explanation and the Causal Structure of the World.* Princeton, NJ: Princeton University Press.

Saussure, Ferdinand de. 1966. *Course in General Linguistics.* New York: McGraw-Hill Book Company.

Smart, J.J.C. 1959. Sensations and Brain Processes. *Philosophical Review,* 68, 141–156.

Sartre, Jean-Paul. 1956. Being and Nothingness. In Robert C. Solomon (Ed.), *Phenomenology and Existentialism* (pp. 466–473). Boulder, CO: Rowman and Littlefield.

Searle, John. 1969. *Speech Acts.* Cambridge, U.K.: Cambridge University Press.

Searle, John. 1979. *Expression and Meaning.* Cambridge, U.K.: Cambridge University Press.

Searle, John. 1983. *Intentionality: an Essay on the Philosophy of Mind.* Cambridge, U.K.: Cambridge University Press.

Searle, John. 1984. *Minds, Brains, and Science.* Cambridge, U.K.: Harvard University Press.

Searle, John. 1992. *The Rediscovery of the Mind.* Cambridge, MA: The MIT Press.

Sellars, Wilfred. 1963. *Science, Perception and Reality.* New York. Routledge & Kegan Paul.

Scheffler, Israel. 1981. *The Anatomy of Inquiry.* Indianapolis, IN: Hackett.

Schlick, Moritz. 1932. Positivism and Realism. In A.J. Ayer (Ed.), *Logical Positivism,* (pp. 82–107). New York: The Free Press.

Schlick, Moritz. 1934. The Foundation of Knowledge. In A. J. Ayer (Ed.), *Logical Positivism,* (pp. 208–227). New York: The Free Press.

Shope, R.K. 1983. *The Analysis of Knowing.* Princeton, NJ: Princeton University Press.

Sklar, Lawrence. 1974. *Space, Time, and Spacetime.* Berkeley, CA: University of California Press.

Soames, Scott. 2002. *Beyond Rigidity: The Unfinished Semantic Agenda of Naming and Necessity.* New York: Oxford University Press.

Soames, Scott. 2003. *Philosophical Analysis in the 20th Century. Volumes 1 and 2.* Princeton, NJ: Princeton University Press.

Soames, Scott. 2005. *Reference and Description.* Princeton, NJ: Princeton University Press.

Smullyan, Arthur. 1948. Modality and description. In L. Linsky (Ed.), *Reference and Modality* (pp. 35–43). New York: Oxford University Press.

Smullyan, Raymond. 2001. Gödel's Incompleteness Theorems. In L. Goble (Ed.), *Philosophical Logic* (pp. 72–90). Malden, MA: Blackwell.

Stalnaker, Robert. 1976. Modality. In J. Garfield and M. Kiteley (Eds.), *Meaning and Truth* (pp. 467–477). St. Paul, MN: Paragon Press.

Stalnaker, Robert. 1984. *Inquiry.* Cambridge, MA: MIT Press.

Stalnaker, Robert. 1999. Context and Content. Oxford: Clarendon Press.

Stevenson, C. L. 1937. The Emotive Meaning of Ethical terms. In A. J. Ayer (Ed.), *Logical Positivism,* (pp. 264–281). New York: The Free Press.

Stich, Stephen. 1978. Scientific versus folk psychology. In D. Rosenthal (Ed.), *The Nature of Mind,* (pp. 591–600). New York: Oxford University Press.

Stich, Stephen P. 1983. *From Folk Psychology to Cognitive Science: A Case Against Belief.* Cambridge, MA: The MIT Press.

Stoll, Robert. 1963. *Set Theory and Logic.* New York: Dover.

Strawson, Peter. 1950. On Referring. In A.P. Martinich (Ed.), *The Philosophy of Language, (*pp. 219–234). New York: Oxford University Press.

Strawson, Peter. 1959. *Individuals.* New York: Routledge.

Strawson, Peter. 1969. "Meaning and Truth." In A.P. Martinich (Ed.), *The Philosophy of Language,* (pp. 91–101). Oxford: Oxford University Press.

Stroud, Barry. 2001. *The Quest for Reality.* Oxford: Oxford University Press. Watson, John Broadus. 1924. *Behaviorism.* New York: Norton.

Wittgenstein, Ludwig. 1922. *Tractatus Logico-philosophic us.* London: Routledge, Kegan & Paul.

Wittgenstein, Ludwig. 1958. *The Philosophical Investigations.* Oxford: Blackwell.

Taylor, Kenneth. 1998. *Truth & Meaning.* Oxford: Blackwell.

Van Fraassen. 1980. *The Scientific Image.* New York: Oxford.

Van Fraassen. 1989. *Laws and Symmetry.* New York: Oxford.

[1]. *"Mutatis mutandis"* means: *provided that the needed changes are made.* Consider the statement:

(i) "if I make sure that I exercise and get plenty of rest, and everybody else does the same thing, the world will be a better place."

(i) is ambiguous. It could mean either:

(ii) "if I do my best to make sure that I exercise and get plenty of rest, and everybody makes sure that I exercise and get plenty of rest, the world will be a better place,'

or

(iii) "if I do my best to make sure that I exercise and get plenty of rest, and each other person makes sure that he or she exercises and gets plenty of rest, the world will be a better place."

Unlike (i), the following sentence *unambiguously* has (iii) for its meaning:

(iv) "if I make sure that I exercise and get plenty of rest, and everybody else does the same thing *mutatis mutandis,* the world will be a better place."

Sentences that are ambiguous in much the same way as (i) often occur in analytic philosophy, and it's often unclear how they are to be disambiguated. But the "mutatis mutandis" makes it easy to disambiguate them.

[2]. Expressions that are italicized and boldfaced are very important ones whose meanings must be known. In most cases, those terms are defined and explained in the analytic index, where they are alphabetically listed. When such an expression isn't defined in the analytic index, it's defined on the *first* page where it occurs. In cases where I judge the expression to be an especially important one, it is defined both on the first page of its occurrence and in the analytic index.

An idea is" "incoherent" if it is expressed by a statement that either explicitly or implicitly contradicts itself.

[3]. Creating the right mechanisms may involve an enormous amount of theoretical knowledge—a N.A.S.A. engineer may know a lot more about

physical law than many a professor of theoretical physics. Whether one is an engineer as opposed to a physicist is a function, not of how much one knows about physics, but about one's attitude towards that knowledge. The engineer is interested in physical law only to the extent that it helps build bridges and tanks. The physicist is interested in bridges and tanks only to the extent that it helps identify physical laws. That is why, even though Newton was one of the greatest physicists of all time, many a contemporary engineer knows more about physical law than he did.

[4]. Dummett (1978, Chapter 25).

[5]. This is now contested. Some philosophers hold that there are coherent statements—statements that don't embody internal inconsistencies—that cannot possibly be true. See Putnam (1975) and Kripke (1980). In Chapter 8, it is argued that this is false.

[6]. "FL" is short for "Frege's legacy."

[7]. See Frege (1980). Russell (1905, 1914, 1917, 1918, 1927, 1948) did a very good job of clarifying and developing this viewpoint.

[8]. Russell (1903).

[9]. The argument about to be given was put forth by Russell (1920).

[10]. See Pap (1949). Also see Carnap (1932) and Schlick (1934).

[11]. "TP" is short for" "tall person."

[12]. Frege used the word "concept," not" "property."

[13]. "LP" is short for "Larry perception."

[14]. "LS" is short for "Larry sentence."

[15]. An expression is" "semantically simple" if it doesn't consist of other expressions. So" "the person in the cockpit at the time of the call" is semantically complex, but" "Socrates" and" "Larry" are not.

[16]. See Russell (1917, Chapter X), Barwise (1983), Kuczynski (2007).

[17]. See Brentano's *Psychology from an Empirical Standpoint.*

[18]. For points similar to those just made, see Russell (1917, Chapter X), and Russell (1984).

Nathan Salmon holds that Bart Simpson, Zeus, etc. do exist. He holds that, in act, square circles and even primes greater than two exist. See Salmon (2005, Chapters 1–3). Salmon's view and his argument for it are discussed in Chapters 6, 8, and 9. It will be seen that they coincide very nearly with those of Brentano and Meinong.

In fairness to Salmon, whose important work I urge all aspiring philosophers to study, I should point out that my criticisms of his viewpoints are heavily biased by my conviction that philosophical problems ought to be adjudicated through use of post-Fregean methods. For this reason, those criticisms should be taken *cum grano salis.*

And there's no denying that Salmon's views have a certain consonance with common sense. Many of my students say that truth is subjective—that everyone has their own truth. If that position is right, then Bart Simpson and Fred Flintstone *do* exist. Their sentiments, I would guess, are more in alignment with Salmon's views than they are with mine.

Salmon also holds that there exist non-existent objects. (See Salmon 2005, Chapters 1–3.) For example, he holds that the sentence "Socrates was wise" is meaningful and, therefore, that there exists something that is its meaning. But he also holds that this very meaning doesn't exist. And Salmon holds that one can rationally believe a proposition *and* its negation—that, for example, one can, without being irrational, believe *Smith is in Virginia* and *Smith is not in Virginia (since Smith is on the Moon*) at the very same time. (See Salmon 2007, especially the chapter titled "Irrational belief.') In Chapters 6, 8, and 9 we'll discuss Salmon's arguments for these views. But it's a matter of secondary importance whether these positions of his are accurate. For should they turn out to be false, they would, for that very reason, be true! Remember what Salmon is arguing for: (i) the non-existent exists and (ii) it is sometimes rational believe P and not-P at the same time. The very fact of there being no good reasons to accept (i) would entail that there *did* exist good reasons to accept it; and the very fact of our having reason to reject (ii) would mean that, at least potentially, we had good reason to accept it. Thus, Salmon's arguments, if unsuccessful, are successful for that very reason, and they are therefore self-validating. One can only marvel at such intellectual virtuosity.

But even if, by some remote chance, the non-existent *does* turn out to be non-existent, and even if self-contradictory beliefs *do* turn out to be unreasonable ones—even if, I dare say, these strictly theoretical possibilities should materialize, the very fact of Salmon's having said otherwise is what is important. In saying that the non-existent exists and that the contradictory can be rationally accepted, Salmon is, it seems to me, trying to break down rigid categories that hobble thought. His fascinating and daring views are like Zen koans, the purpose of which isn't so much to show one the truth, or to delineate truth-conducive methods of reasoning, as it is to rethink one's views as to what it is that one is learning in learning the truth.

[19]. This apt way of summing up special relativity is due to Howard Gardner (1976).

[20]. Empirical discoveries in psychology and physics do play a part in the dissolution of philosophical problems. But usually when some empirical discovery sheds light on a philosophical issue, it is by prompting philosophers to produce the relevant analytic knowledge. Certainly a lot of good philosophy was *prompted* by what Einstein and Freud had to say. But the ensuing philosophical insights were ones that, theoretically, could have been arrived at independently of what they said.

[21]. And Frege's knowledge of the meanings of specific German sentences, while obviously empirical, wasn't materially implicated in the points he was making. He could have made those points had he not known a word of German and had been writing in, for example, Hindi or Spanish.

[22]. Wittgenstein (1958).

[23]. Wittgenstein (1922).

[24]. Epistemology is the theory of knowledge.

[25]. See Fodor's (1990) book *A Theory of Content* and his (1998) book *Concepts.*

[26]. See my (2007) book *Conceptual Atomism and the Computational Theory of Mind.*

[27]. Although this position is not as popular as Fodor's, I am not alone in holding it. For example, Laurence Bonjour (1998) holds it.

[28]. Actually, Quine, in his paper "Epistemology Naturalized," said that epistemological questions are not to be answered in this way, and are to be answered in the same way as scientific questions; e.g., "what is the Moon's chemical composition?" There are two points to make here. First of all, even though Quine was an analytic philosopher by trade, it doesn't follow that all of his views were those of an analytic philosopher; and in this context, I would suggest, he was not being an analytic philosopher. He was being an opponent of analytic philosophy. Second, Quine's view is incoherent. Natural science is based on empirical observation. Empirical observation cannot tell you that it itself is a source of knowledge. Therefore, it cannot be known through natural science that observation is a source of knowledge, and it must there-fore be known in some non-empirical, and therefore non-scientific, manner. Jaegwon Kim (1993) makes a similar point.

[29]. Incidentally, the very fact that JK is to be adjudicated in a non-observational manner shows that JK is false. Supposing JK correct, the fact it describes is that one statement is inconsistent with another; and that fact is obviously causally inert, that being why it cannot be observed.

[30]. In this context, the word "sentence" will stand for both ***sentence-types*** and ***sentence-tokens;*** i.e., for sentences *per se* and also for people's utterances and inscriptions of them.

[31]. See Frege (1918).

[32]. In the sentence, "x is a certain way,' the word "is" doesn't denote the relation of identity. The "is" is being used ***predicatively.*** It's being used in the way it is in the sentence "Smith is happy,' which obviously doesn't say that Smith is *identical* with happiness, and not in the way it's being used in the sentence "Smith is the guy who stole my car." See Chapter 16, Section 1.1 for further discussion.

[33]. Actually, we'll see in Chapter 2, that propositions are *sets* of properties. But this is a nuance. And it's an irrelevant nuance, since, as we'll see in Chapter 2, sets *are* properties and, therefore, propositions are identical with sets. This must be understood aright. The property of being a bird is *not* identical with the set of all birds—contrary to what Quine (1960, 1970) and other anti-Platonist philosophers have held.

[34]. And so far as it *is* up to us how it is, it's up to us in a *causal* sense. It is people's actions, not their opinions, that make things true. Even though, on the basis of an opinion, one may perform acts that wrench the world into a certain form, one's opinion *per se* is innocuous: the world is the way it is regardless of that opinion.

Also, it's not a matter of opinion what opinions we have; it's a matter of fact. We have the opinions that we have and, so far as we know ourselves, our opinions as to what those opinions are track those pre-existing facts. Our second-order opinions—i.e., our opinions about opinions—no more make our first opinions be the way they are than our first-order opinions about the stars make the stars be the way they are. Our second-order opinions may induce us to engage in acts that result in the modification of our first opinions, whereas, our opinions about the stars cannot possibly induce us to engage in acts that change the stars. But that's irrelevant, since it doesn't phase the fact that it is never *in virtue of* what one's opinions are that the world is, or is not, in any respect, a given way.

[35]. This is the branch of philosophy that aspires to clarify the concepts that are used in formal logic but that themselves resist formal treatment; e.g. ***entailment, formal truth, informal analytic truth, possibility, logical form,*** etc. See Section 7.0 for further discussion. Most of what is said in Part 1 of the present book falls within the jurisdiction of philosophical logic.

[36]. At certain junctures, this belief of theirs is correct—but only to a certain degree. Logic and rigor only go so far. They're no substitute for insight or creativity—especially not in fields such as ethics, politics, and religion. Logic and rigor have only a secondary function. They're helpful in evaluating insights already generated. But they're useless when it comes to generating them. Political philosophy has benefited enormously from post-Fregean love of rigor. For some reason, political philosophers didn't take it too far: respect for rigor did not, in their case, turn into rigor-worship. The same cannot be said of every other branch of analytic philosophy. As we'll see in Chapter 25, the philosophy of religion has become one of the most arcane and technical branches of philosophy, and it's not always clear that what analytic philosophers of religion are doing has much to do with religion.

[37]. “Acoustical” means “noise-related,” and “morphological” means “shape-related.”

[38]. But, as a matter of exegetical fact, I have in no way exaggerated Wittgenstein’s claims.

[39]. He hedges by “roughly speaking, meaning is use.” He never explains the hedge. I’ve omitted the first two words, since they serve no identifiable purpose.

[40]. “WA” is short for “Wittgenstein’s argument.” This is not a quotation. There is no one passage where Wittgenstein argues on behalf of his thesis that meaning is use. His arguments are dispersed over many pages, in most cases being presented in the form of aphorisms and rhetorical questions. Quine (1960) advocates the position about to be presented, and he states it very crisply. Some of the locutions that occur in WA are found in Quine (*op. cit.*), and some, though not (so far as I know) used by Quine himself, are Quinean.

[41]. Wittgenstein is right to assume that propositions, if they exist, are non-spatiotemporal. (What follows is my own defense of this assumption.) First of all, propositions, if existent, aren’t facts. “JMK owns a hundred yachts” is meaningful; if its meaning were the *fact* that JMK owns a hundred yachts, it would have no meaning, since there is no such fact. Propositions obviously aren’t rocks or trees or explosions or anything that we’ll encounter in the external world.

Nor are they mental entities. *The* proposition that snow is white isn’t some event or structure in my mind or in yours or in anyone else’s. That proposition will exist after we and our ideas are gone. It doesn’t help to say that propositions are artifacts, like churches and lawn-mowers. Churches and lawnmowers are objects in the external world; propositions are not.

I’ve heard it said that propositions, though not constituents of this or that person’s mind, are constituents of some “collective mind.” But so far as it has any meaning, this proposal is useless. If, in saying that propositions are ideas in a “collective mind,” one is saying that they are ideas in the minds of various individuals, one is simply wrong as we just saw. If, in saying this, one is saying that they are ideas in some mind that, though

composed of various individuals mind, isn't identical with any one of them, one is making the doubtful, and very likely incoherent, assumption that there exists such a mind. We may conclude that no proposition is identical with *anything* you'll find if you take an inventory of the spatiotemporal world. If they exist, propositions are non-spatiotemporal entities.

[42]. The argument being put forth is found in precisely this form in Quine (1960). It is also Wittgenstein's position. But Wittgenstein doesn't state it as crisply as Quine.

[43]. A person who *did* know that language would agree that I hadn't said anything *after* that person had learned that I didn't know what they meant. That's why if those noises were offensive, that person would cease to be offended *after* he learned that I didn't know what they meant.

[44]. There are different views as to what it is. But, as is universally conceded, they are not themselves sentences. See Chapter 4.

[45]. 'TA' is short for 'Tractarian argument.' TA is a paraphrase, not a quotation. But some parts (e.g., BG) are direct quotations.

[46]. *Tractatus Logico*-Philosophicus, Section 4.003. I am using the Pears/McGuinness translation.

[47]. *Tractatus* Logico-*Philosophicus,* Section 4.003. I am using the Pears/McGuinness translation.

[48]. Ryle (1949).

50. Wittgentein acknowledges that his view has this consequence. At the very end of the TLP, he says that anyone who understands what he has said will see that it is meaningless. But this is incoherent: to understand a statement is to know what its meaning is; so Wittgenstein's words can't be understood unless they are meaningful.

51. Russell (1905) holds (correctly, I will argue) that one cannot possibly say of some individual x either that x does exist or that x does not exist. When one says "the even prime greater than two doesn't exist," one isn't saying *of* some number x such that x is an even prime greater than two that x doesn't exist. If *that's* what one were saying, the existence of just such a number would be presupposed by that very statement. What one is

actually saying is that the class of even numbers greater than two doesn't overlap with the class of prime numbers. All meaningful **negative existentials** are to be similarly analyzed. (Incidentally, Nathan Salmon (2005, Chapters 1–3) disagrees with this. He says that one *can* both meaningfully *and correctly* say of some object x that x doesn't exist. He says, in other words, that it's possible for there to exist an object x such that one meaningfully and correctly says that x doesn't exist. This is, quite obviously, false. In Chapters 6, 7, and 18, the problems with Salmon's view are detailed.)

Supposing that Russell is right about this, it might follow that NS_2 *is* meaningless, since NS_2 seems to say of some individual x that x doesn't exist. Because I believe that Russell's analysis of negative existentials is correct, I agree that *if* NS_2 is indeed meaningless *if* it says of some individual x that x doesn't exist. But it's pretty clear that those who say that NS entails NS_1, and who therefore implicitly accept NS_2, are saying *something.* The problem isn't that their words lack meaning, but that they bear a very wrong meaning. Since NS_2 *is* meaningful, and since it would be meaningless if it said *of* some individual x that x didn't exist, it must be taken to mean that:

(NS_3) "there exists a square circle, and there fails to exist anything that is both square and a circle," which, though absurd, is meaningful.

52. A conjunct is one of the sentences composing a conjunction. A conjunction is a sentence that consists of two or more sentences that are joined by an "and" (or some other comparable expression, e.g. "but").

53. See Appendix 2 for a discussion of this principle.

54. "CT" is short for "central thesis."

55. "Corollary" is a synonym of "consequence." A corollary of nation X's having all its troops in Nation Y is that it can't station any troops in Nation Z.

56. See Wittgenstein (1922), Carnap (1932, 1934), Schlick (1932, 1934), Hahn (1933), Stevenson (1937), Hempel (1950), Ayer (1952).

57. "Iff" means "if and only if." "S_1 iff S_2" means "if S_1 is true, it follows

that S_2 is true; and if S_2 is true, it follows that S_1 is true." Thus, "S_1 iff S_2" means that the truth of S_1 is both ***necessary*** and ***sufficient*** for the truth of S_2.

58. In addition to being false, the statement:

(VE) "all meaningful non-tautologies are capable of being observationally verified"

isn't a tautology, since it's debatable (and false, as we saw). Given that VE isn't a tautology, one who accepts VE must assume it verifiable if not meaningless, and must therefore assume it verifiable. But VE is unverifiable for the same reason as "all metal expands when heated." Any attempt to show otherwise would be self-refuting. One can't try to show that VE is verifiable without first assuming it meaningful. But in assuming this, one is assuming that unverifiable non-tautologies *can* be meaningful and, therefore, that VE is false.

59. The term "falsificationism" is associated with Karl Popper (1902–1994). But, contrary to what is sometimes alleged, Popper didn't advocate the doctrine to which that expression refers in this context. He did not hold that a statement had to be falsifiable to be *meaningful.* Popper accepted the doctrine that for a statement to be *scientific* is for it to be falsifiable. (Both doctrines are known as "falsificationism." Hence the mix-up.) He makes this clear in his (1983) book *Realism and the Aim of Science.*

60. Carl Hempel (1905–1997) made points similar to the ones just stated. (See Hempel 1965, Chapter 3.) Hempel starts out by identifying two principles. (i) The ***negation*** of a meaningful statement is itself meaningful. (ii) A meaningful statement can neither entail nor confirm a meaningless one.

The negation of "snow is white" is "it is not the case that snow is white" or, more simply, "snow is not white." For one sentence to negate another is for the one to affirm the falsity of the statement made by the other. If a sentence has no meaning, and thus makes no statement, there is no statement whose falsity its negation affirms and, consequently, no statement that it makes. (Given that 'arga bunga blurb' makes no statement, there is no statement whose falsity is affirmed by "it is not the

case arga bunga blurb.") Hence (i).

For one sentence to entail another is for the statement made by the second to be a consequence of the statement made by the first. Thus, "x is a triangle" entails "x has more than one side," since the statement made by the second follows from the statement made by the first. Thus, no sentence that has a meaning (i.e. makes a statement) can entail one that does not. It follows that no sentence that has a meaning can confirm one that does not. Hence (ii).

Some verifiable sentences (e.g., "some metal object doesn't expand when heated") have unverifiable negations ("all metal objects expand when heated"). Thus, verificationism is false. Some falsifiable sentences (e.g., "no metal objects expand when heated") have unfalsifiable negations ('some metal object expands when heated'). Thus, falsificationism is false. Some meaningless sentences (e.g., 'either the universe is a perfect unity or snow is white') follow from confirmable ones (e.g., 'snow is white'). Thus, confirmationalism is false.

According to some philosophers, Hempel points out, a sentence S is meaningful if it can be translated into a statement that concerns only one's sensations. (George Berkeley[60] (1685–1723) held this, as did Ernst Mach[60] (1838–1916) and Rudolph Carnap[60] (1890–1970).) This contention is false, Hempel rightly points out, given that it's a very strong version of the view that only completely verifiable sentences are meaningful.

The fact that this view is false shows that ***strict empiricism*** is false. Strict empiricism is the view that all non-trivial knowledge has a strictly sensory basis. To say that knowledge has a strictly sensory basis is to say that what is known is either (i) directly observed to be the case or (ii) is inferred from what is directly observed to be the case by means of an inference-rule of whose legitimacy one's own observations made one aware.

In Section 7.1, we'll consider another, subtler argument of Hempel's for the position that not all knowledge is observation-based.

61. In Section 4.8 we'll question this assumption.

62. Technically, ink marks, noises, etc. are sentence tokens, not sentences. But we can set that aside here.

63. Carnap (1934).

64. He didn't put it this starkly. And he probably wouldn't have accepted it—at least not for long—if it had been presened to him this starkly. But that doesn't mean that it wasn't his view. It was.

65. This is the view of Wittgenstein (1958), W.V.O. Quine (1960), Hartry Field (1977), Robert Brandom (1994), and John McDowell (1998).

66. "NT" is short for "Newton's theory."

67. Note: the bracket-like expressions that occur in what follows—the "‹" and the "›"—are known as ***quasi-quotation*** marks. They occur frequently in words on the philosophy of language and they will occur frequently in this book. In this context, they may be treated as quotation marks. But there are subtle differences between them and quotation-marks, because of which they are not technically interchangeable with quotation-marks. These are identified in Appendix 2, where a definition of quasi-quotation is given. See the entry titled "quasi-quotation marks.

Quine (1940) invented quasi-quotation marks. In so doing, he did a great service to semantics and logic, for reasons that are given in Appendix 2.

68. Strictly speaking, it doesn't quite say this. S would say how a bodies x_1 and x_2, having masses m_1 and m_2, would behave *to the extent that* their behaviors would be functions of their having those masses and being that distance from one another. But other factors might obviously be involved; e.g., x_2 might be near some third body x_3 whose gravitational pull was much stronger than x_1s' or it under the influence of some non-gravitational force.

69. See Hempel (1965, Chapter 3), Sheffler (1981, Part II).

70. Wittgenstein, himself LP's founding father, *did* try to make a case that LP was fundamentally misguided, and not just technically faulty. He spent 20 years trying to make a case for this, and his efforts were embodied in a now famous book known as the *Logical Investigations.* But the doctrines put forth in the *Investigations* are quite as incoherent as LP itself. See

Chapter 5.

71. Here is a similar argument. LP is a theory—a theory concerning the conditions sentences must satisfy to be meaningful. Tautologies aren't theories. Tautologies are conventions. Conventions aren't theories. Therefore, LP isn't a tautology.

72. I'm quoting from the McGuiness/Pears translation.

73. Russell (1941, pp. 14–15).

74. In her very good book *Physics and the Philosophers,* L. Susan Stebbing makes similar points. (Stebbing is discussing an argument given by Arthur Edington that is similar to Russell's.)

75. The argument about to be presented is further discussed in Chapter 18.

76. "GA" stands for "Goodman's argument." This is not a quotation.

77. What follows wasn't said by Goodman; I'm developing his point.

78. See Mates (1952) for some arguments that, in my judgment, embody a failure to take these points into account.

79. Wittgenstein says that the alphabet is a picture 'in the ordinary sense' of the corresponding sounds—that "k" is a picture, "in the ordinary sense," of the corresponding sound. That isn't true. The relationship between "k" and the corresponding sound is conventional, not iconic.

80. A semantic rule is what is expressed by any sentence saying what is meant by an expression of a language.

So if, while pointing to individual x—who, we will stipulate, is Mick Jagger—I say "that guy's name is Mick Jagger," I'm stating a semantic rule, since what I'm saying is that "Mick Jagger" refers to x.

81. Frege, who held the same view, said that "propositions are true facts." Frege's view isn't quite accurate. False propositions aren't facts. It isn't a fact *that the moon is made of cheese.* Bearng in mind that false propositions aren't facts, let's suppose, with Frege, that true propositions *are* facts. Facts are, in at least some cases, spatiotemporal entities. The fact of my typing right now consists of various displacements of mass-energy.

So if true propositions were facts, then some propositions would be spatiotemporal entities, whereas false ones would categorically fail to be such entities. But it would be theoretical arbitrariness of the worst kind to hold that some members of the class of propositions—which is obviously a unified, highly integrated class—were spatiotemoral, whereas other members of that same class were not.

82. When putting forth the picture theory of meaning, Wittgenstein uses the German word "satz," which means "sentence." But Wittgenstein's translators sometimes translate it as "proposition." Supposing that they're right to do so, the picture theory of meaning is the view that propositions, not the sentences that express them, are pictures of the facts they describe.

Thus interpreted, Wittgenstein's thesis is false. Like the states of affairs they depict, actual pictures (e.g., drawings, photographs) are spatiotemporal entities, and can thus bear a *physical* resemblance to those states of affairs. But propositions are not spatiotemporal entities; they're abstract objects. Therefore, they cannot in any literal sense resemble the facts to which, when they're true, they correspond. To be sure, propositions must resemble those facts in some way or other. But that's trivial, given that, for any two things, there is a respect in which they're similar to each other.

A terminological point before we continue: Given a proposition, we'll refer to the fact that must hold for that proposition to be true as its "truth-maker." So, supposing that John is North Carolina, the truth-maker of that proposition is the fact that John is North Carolina.

For the theory that proposition are pictures of their truth-makers to be have any substance, it would have to be said how exactly they resemble those truth-makers; and in order to do that, it would have to be said what exactly propositions are. Usually, Wittgenstein spurns propositions. But every now and then he seems to concede their existence. So far as he does so, his view is that they are veritable models of their truth-makers—that, in other words, a given proposition actually consists of the very objects involved in the fact that, supposing it true, it describes and that, in that proposition, those objects are interrelated in the way they would be in the corresponding fact, were the latter to exist.[82] But that theory is rank

absurdity. For, if true, the proposition can't even exist without being true. If the proposition that Smith punched Jones is a structure consisting of Smith and Jones in which the former punches the latter, then that proposition is *it self* a case of Smith's punching Jones. And in that case, that proposition can't even exist without being true. But that proposition can exist without being true. The proposition that JMK jumped over the Empire State building exists but isn't true.

There's also the problem that, even though Socrates no longer exists, the proposition *Socrates admired Pal to* still does. Interestingly, Nathan Salmon (1986, 2005, 2007) holds that, for this very reason, that proposition doesn't exist. Salmon's position is that the sentence "Socrates admired Plato" is meaningful and that it's meaning is the proposition that Socrates admired Plato. But, says Salmon, that meaning is non-existent, the reason being that Socrates, one of components, is non-existent. Thus, says Salmon, there exists a proposition that doesn't exist that is the meaning of "Socrates admired Plato." This is what Salmon *himself* says. It is not an unintended consequence of his view; it *is* his view. (See Salmon (2005), Chapters 2 and 3.) The merits of this view are discussed in Chapter 3.

In any case, Salmon's view is a quasi-Tractarian one. He says that (so far as it exists) the proposition that Socrates admired Plato consists of Socrates and Plato—those two individuals themselves, not concepts thereof—and that, in that proposition, the former stands in the relation of admiring with respect to the latter or stands in some analogous relationship to him. Apart from the fact that Socrates and Plato aren't around to do the work that Salmon's theory requires of them, that theory has the drawback that the mere existence of the proposition in question prejudges a number of open empirical issues. If, in that proposition, Socrates must *actually* admire Plato, then the mere existence of that proposition predetermines its own truth. If it's some other relationship that Socrates bears with respect to Plato, some other issue will be prejudged. In addition, the question will arise why it is that the proposition in question consists of Socrates' bearing *that* relation to Plato. What is so special about that relationship? And to answer this question, Salmon would actually have to identify the relationship in question.

To Salmon's credit, he has attempted to identify this relationship. He has said that the proposition that Socrates admired Plato is the following ordered pair: <the relationship of admiring <Socrates, Plato>>. Thus, he has said, in effect, that the relationship in question—the one that Socrates bears to Plato in the proposition in question—is one that a thing x bears to a thing y in virtue of its being the case that the following ordered pair exists: <the relationship of admiring, <x, y>>.

This is certainly very helpful. But we have a slight problem. Ordered pairs aren't true or false. The ordered pair <the relationship of admiring <Socrates, Plato>> true or false. (John Perry points makes this exact point.) An alternative to Salmon's theory is put forth in Chapter 3.

83. Goodman (1976) discusses the concept of a digital representation as well as the, soon to be mentioned, concept of an analogue representation.

84. The differences between iconic and linguistic representation are discussed at length in Chapters 21–23 of my (2007) book *Conceptual Atomism and the Computational Theory of Mind: A Defense of Content-internalism and Semantic externalism.*

85. "Grass is greed and snow is white" is the result of ***conjoining*** "snow is white" and "grass is green," and "Grass is greed or snow is white" is the result of ***disjoining*** them.

86. A statement is vacuously true if it's trivially incapable being counter-exampled. An example is: "every human being who lives on Neptune plays the accordion" In order for that sentence to be false, there would have to be a human being on Neptune who didn't play the accordion. Since there are no human beings on Neptune to begin with, there are none who don't play the accordion.

87. Berkeley made this point very clearly in his (1734) work *The Principles of Human Knowledge.*

88. Sometimes the word "'iconic" is used to mean a picture-*like* linguistic representation that isn't really a picture. Thus, hieroglphys are "iconic," since they look like the things they represent. But in this context, 'iconic' will be synonymous with 'pictorial', 'imagistic,' etc.

89. Short for "image of a triangle."

90. This is an adaptation of the Pears/McGuinness translation. I've replaced each occurrence therein of "proposition" with "sentence," for the reasons given earlier. The italics are in the original.

91. I suspect that it was his awareness of this fact that led him to end his book by saying that everything in it is nonsense.

92. Wittgenstein's (personal and philosophical) mentor, Bertrand Russell, was at that point the world's authority on paradoxes of this kind. Russell's work on these puzzles is by no means obsolete. Wittgenstein was in daily contact with Russell during much of the five year period preceding his writing the TLP.

93. And make them go away is what, in effect, Russell (1908) tried to do. Russell said that, even though they seem meaningful, statements that, like (i), self-refer are meaningless. They're pseudo-statements. They have the surface structure of real statements, but not the logical form.

Russell found that some intuitively reasonable logical principles entail self-contradictory statements and must therefore be false, and he found that those principles *in fact* referred to themselves, even though they didn't seem to. The most famous example concerned the ***axiom of comprehension,*** viz: (AC) Given any property, there is some class of objects containing everything that has that property and nothing that lacks it.

AC is obviously a very reasonable principle. Consider the property of being a bird. There is some class that contains every bird in existence, and doesn't contain any non-birds.[93] But Russell poked a major hole in AC. Let K be the class of all classes that don't belong to themselves. K must belong to itself if it doesn't, since it's the class to which all classes that don't belong to themselves belong. But, for the same reason, K cannot belong to itself if it does. But, for the very same reason, if K belongs to itself, then it doesn't belong to itself. So K belongs to itself if it doesn't, and it doesn't if it does. It follows that (K) the class of all classes that don't belong to themselves is a member of itself

entails its own negation, and that the negation of K entails *t i s* own

negation. Since, according to the law of excluded middle (*any given proposition is either true or false*), either K or its negation is true, it follows that K *and* its negation are true. But, according to the law of non-contradiction (*no proposition s i both true and false*), this isn't possible.

Russell dealt with this by saying that nothing can be true of itself. Properties cannot be their own instances; statements cannot describe themselves. Thus, AC is ruled out, and so is (i).

94. A historical point: In his (1734) book *Principles of Human Knowledge,* George Berkeley made virtually every point that Wittgenstein made in the *Tractatus.* Moreover, in the *Principles* and in other works (e.g., *An Essay Towards a New Theory of Vision*) Berkeley made a number of cogent, insightful, and strikingly modern points, all of which are absent from the TLP. Berkeley's important work will be discussed in Chapter 13.

The one claim that is clearly present in the TLP that is clearly absent from the *Principles* is the thesis that for a sentence to be meaningless is to be syntactically ill-formed. Berkeley claimed that for a sentence to be meaningless is for it to be unverifiable, i.e., incapable of being definitively shown to be true on the basis only of what one's senses tell one. In other words, Berkeley was a verificationist.

Of course, the TLP advocates verificationism. But verificationism is inconsistent with one of the other main claims of the TLP, namely, that meaninglessness is categorically the result of bad syntax. I leave it to the reader as an exercise to demonstrate this.

95. Such references are found in the work of Russell (1918), Wittgenstein (1922), and Hempel (1950).

96. A "logically perfect language" isn't the same thing as a formal language. But, like many others in his time and ours, Hempel didn't see this. The thesis that meaningful sentences are those that can be expressed in logically perfect languages is, I am certain, an expression of this confusion.

97. Hempel (1952, 1965).

98. "HA" stands for "Hempel's argument."

99. Hempel (1952) does an extremely good job of making it clear what conditions an object must fulfill if it is to serve as a metric standard.

100. To say that $\sqrt{2}$ is an irrational number is to say that there are no whole numbers p and q such that $p/q = \sqrt{2}$. This entails that, given a length L_1 of $\sqrt{2}$ units and also a length L_2 of N units, where N is a rational number, there is no rational number M, such that each of L_1 and L_2 can be divided, without remainder, into M units. Two lengths that are related to each other in this are "incommensurable." It was, it is said, Pythagoras who discovered the existence of incommensurables. Richard Dedekind (1831–1916) and Bertrand Russell (1872–1970) jointly provided the first clear analysis of incommensurability. See Chapter 7.

101. As previously stated, a number N is irrational if there are no numbers p and q such that $N = p/q$. See Chapter 7.

102. Hempel (1952, 1965).

103. See Nagel (1962), especially the chapter titled "Indeterminism and causality in quantum-physics." Nagel brilliantly argues that, although sub-atomic phenomena are indeterministic *with respect to* the categories used to described macro-scopic (and otherwise super-atomic) phenomena, it is possible to identify categories with respect to which sub-atomic phenomena are deterministic. Philipp Frank (1949) has a similar view.

104. See van fraassen (1989).

105. The plural of "sentence-schema" is "sentence-schemata."

106. In Chapter 13, it is argued that, in his (1734) work *Principles of Human Knowledge,* George Berkeley stated these very points, or at least came close to doing so.

[49]. Russell makes this very point in *An Inquiry into Meaning and Truth.*

[50]. E.g. Wittgenstein (1922).

[51]. The word "nomic" means "having to do with natural law."

[52]. See Armstrong (1993).

[53]. Quine (1960) held it. (See also his book *The Roots of Reference.*) In his view, the property of being water is identical with the scattered object

consisting of all and only water-molecules. It's now a popular one. I suspect that this is largely because Quine countenanced it.

[54]. Quine (1960, 1981) has this view. So do William of Ockham (see his book *Logical Problems*) and Bertrand Russell (1973).

[55]. Plato himself gives this argument in his dialogue *the Parmenides,* and so did Aristotle in the *Metaphysics.* Aristotle gave it its current name.

[56]. This is pronounced "one to one."

[57]. See Russell (1920).

[58]. See Benacerraf (1965), Dretske (1982), and Fodor (1987, 1990, 1998).

[59]. Locke advocates this view. Many who do not explicitly advocate it are committed to it; in other words, the doctrines they do explicitly advocate are false unless conceptualism is true. Every form of empiricism presupposes the truth of conceptualism.

[60]. This argument is given by George Berkeley (1710).

[61]. Ayer (1952) and Carnap (1934) give this argument, or at least assume its cogency.

[62]. The argument about to be given was put forth by Quine.

[63]. Quine (1966) gives a similar argument.

[64]. C.S. Peirce puts forth a similar argument and was the first person, so far as I know, to use the words "type" and "token" in this way.

[65]. In any case, anything that is in space-time is an actualized possibility. But not all actualized possibilities are in space-time. The number two, for example, is an instance of the property of being even, and is thus an actualization of the possibility of there being even numbers. But the number two is not in space-time.

[66]. This chapter is very hard. During a first reading of this book, only read through Section 2.2.

[67]. For expository simplicity, we're leaving aside niceties relating to tense-markers and place-indicators.

[68]. Brandom (1994) holds this view. See Fodor and Lepore (2002) for

some incisive criticisms of this view. This view, then, is not completely anachronistic.

[69]. Cf. Lewis (1986).

[70]. Cf. Quine (1960) and (1981).

[71]. See Grice's classic (1957) article "Meaning."

[72]. In this context, the word "people" is meant to refer to sentient creatures of any kind—Martians, dolphins, etc. I wish to leave it open whether non-human creatures have conventions or, therefore, languages.

[73]. I say "typically" because not all linguists or philosophers of language consider this division to be a legitimate one. Wittgenstein (1958), for example, thought that semantics collapsed into pragmatics.

[74]. The "R" stands for "right," as in "the right analysis."

[75]. "CA" stands for "complete analysis."

[76]. Given that I'm only indicating in broad terms what the discipline of semantics is about, I've omitted some steps from this argument. See Kaplan (1989) for a more thorough analysis of sentence-utterances like the one being discussed (viz. sentence-utterances that contain ***demonstratives***—words like "this" and "that.") There's a second reason why my adding the missing steps unnecessary. The point I'm trying to make (viz. that it isn't always obvious what is literally meant by an expression) is *itself* illustrated by the fact that it may be debated whether U1 and U2 have the same meaning.

[77]. Although he does an, in some ways, creditable job of defending this analysis of utterances containing demonstratives, Kaplan doesn't answer this very question. And this question, it seems to me, is much more important than the narrowly semantic question of what the literal meanings of utterances of "that man," "this world," etc., are. Kaplan notes the difficult (viz. that sentence-tokens that have the very same literal meaning can differ dramatically in what they communicate), and responds by saying (I quote, without omitting relevant contextual material): "Uh oh!" (See Kaplan (1989).)Which pretty much exhausts what he has to say about the matter. It also exhausts what Kaplan-disciples, such as Nathan

Salmon, have to say about the matter. (See Salmon (1986), (2005), (2007).)

[78]. In 2004, in the journal *History and Philosophy of Logic.*

[79]. Note: To understand this section, it is necessary to know what the words "token" and "type" mean when used in semantic discourse.

[80]. Lewis (1975).

[81]. See Blackburn (1984).

[82]. One of the problems with logical positivism was that it assumed, incoherently, that *expressions* could be meaningless. There can be meaningless expression-like entities—but no meaningless expressions. And if, as the logical positivists contended, "there is no free will in a deterministic universe" is truly meaningless, then it isn't a sentence at all; it is, at best, a sentence-like entity. But it plainly is a sentence. It may be (though I do not myself believe this) that "there is no free will in a deterministic universe" is in some way or other an *incoherent* statement. But an incoherent sentence is one that bears a self-undermining meaning. And there's a big difference between bearing a self-undermining meaning and bearing *no* meaning. The logical positivists didn't make this distinction.

[83]. Jerry Fodor and Ernie Lepore do a beautiful job of stating this argument and others like it in *The compositionality Papers.*

[84]. See Kaplan (1989) for points very similar to those given in Sections 7.1–7.21.

[85]. Mediating concepts are obviously similar to what Frege refers to as 'senses.' See Chapter 5. But, not wanting to prejudge the delicate question of whether they're identical with Fregean senses, I've decided to give them a different name.

[86]. My personal view is that *ultimately* definite descriptions are devices of reference. But they are not to be understood in remotely the way that Frege thought. My official policy in this book is that they are to be understood in the way that Russell proposed. Russell's analysis, though not (in my view) a correct answer to the narrowly *semantic* question 'are definite descriptions referring terms?' makes it much easier to absorb the larger, more important epistemological, logical, and metaphysical points

that it was his real purpose in putting forth the Theory of Descriptions to defend. See Kuczynski (2007), Chapters 1–4 for a defense of the view that definite descriptions are (non-Fregean) referring terms.

[87]. See Barwise and Perry (1983) and Kuczynski (2006).

[88]. "IA" is short for "interior angles."

[89]. "BH" is short for "Bill's holdings."

[90]. Barwise and Perry (1983) make similar points. See Kuczynski (2007) for a thorough defense of this position. P.F. Strawson (1950) was the first author to put forth a viewpoint of this general kind.

[91]. Frege (1954).

[92]. Two statements are compatible if neither entails the negation of the other.

[93]. "WRS" stands for "wrong rule for 'someone.' "

[94]. Arguably, not all sentences containing 'someone' have the form ‹someone has psi.› For example, "Plato kicked someone" seems not to have that form. For this reason, RFS, at least arguably, isn't adequate. A more accurate, because more general, definition of "someone" is:

RFS*) In virtue of having the form ‹...someone...›, a sentence says that the property ...*x*... is instantiated. Thus, "Plato kicked someone" says that the property being a thing that is kicked by Plato is instantiated. RFS* collapses into RFS, the reason being that any sentence of the form ‹...someone...› can be represented as having the form ‹someone has psi.› For example, "Plato kicked someone" says the same thing as "someone is such that Plato kicked that person."

What we just said about 'someone' is true *mutatis mutandis* of all other expressions that are to be defined contextually, which, given the fact that all expressions are to be defined contextually, means that it's true of all expressions.

[95]. 'PS' stands for "the property of snoring."

[96]. Frege (1891, 1892b). A research program in linguistics, known as "Montague grammar" is based on developments of these points of Frege's due to Richard Montague (1930–1971). See Montague (1974).

[97]. Wittgenstein (1958) and new-Wittgensteinians, such as John McDowell, hold this. See

McDowell (1994, 1998). Donald Davidson has argued for this thesis on several occasions. See his book *Inquiries into Truth and Intepretation.*

[98]. The most well-known advocate of this view is Jerry Fodor. See Fodor (1975, 1981, 1987). See also Fodor and Pylyshin (1987) and Pylyshin (1984).

[99]. Fodor (1998b) makes similar points (though not quite coincident ones).

[100]. Fodor (1998b) gives arguments similar to, if not coincident with, this argument and the next.

[101]. Frege (1918) said that, in all likelihood, *all* human beings cannot reason without the help of mental imagery.

[102]. There isn't anything it is like to believe that 2 + 2 = 4, but there is something it is like to experience a pain or a pleasure or a tickle. The latter, but not the former, thus have "phenomenology"—they are "phenomenologically pregnant" or, put another way, have "phenomenal content." Anything that there is anything it is like to experience *ipso facto* has phenomenal content (i.e., is phenomenologically pregnant).

[103]. Robert Brandom (1994) accepts Wittgenstein's position. And, on the basis of it, he arrives at the conclusion that individuals *never* think—that it is only whole societies that do so. Brandom doesn't consider any of the points made in Section 6.0.

[104]. "Quale" (plural, "qualia") is a sensation or, more generally, anything that is felt.

[105]. Hume (1740, p. 252).

10. See Freud (1915) for a similar viewpoint.

11. See Freud (1923) for similar remarks.

12. He does this in Wittgenstein (1958).

13. I say "seemed" because his writing is very obscure. And no sooner does somebody claim to have interpreted Wittgenstein than somebody else, often a Wittgenstein expert, claims that Wittgenstein was in fact saying the exact opposite of what the first person claimed. For example, Saul Kripke (1983) wrote a very helpful book in which he put forth plausible interpretations of two very famous argument's of Wittgenstein's—the very one's that we'll consider in this section. Kripke's interpretation fits the text pretty well. But after he'd published this book, two Wittgenstein experts from Oxford University claimed that, although Wittgenstein had said those things, he was being ironic when doing so. He was, they said, saying the exact opposite of what he seemed to be saying and of what Kripke said that he was saying. I'm no Wittgenstein expert. But it does seem a little odd that, in a three hundred page book, Wittgenstein wouldn't anywhere have said that he was being ironic. In any case, the interpretations that I'll put forth are vaguely similar to Kripke's.

14. "RF" stands for "rule following."

15. Like this argument of Wittgenstein's, David Hume's argument for his analysis of personal identity assumes that there is nothing more to a person's mind than the images and sensations that populate his consciousness. We'll consider Hume's analysis, and his argument for it, in Chapter 16. But Hume's argument doesn't depend as deeply on this assumption as does Wittgenstein's. There is an argument very similar to Hume's, for a position very similar to the one it's supposed to

establish, that doesn't involve that doubtful assumption. But nothing even remotely like Wittgenstein's argument is free of that assumption, and there is not, as far as I can tell, any interpretation of Wittgenstein's argument, no matter how liberal, whereby it comes close to making a case for anything at all plausible, let alone accurate.

16. **And it's thus—so Wittgenstein thinks—a case of the "homunculus fallacy."**[16] See Searle (1992). Searle regards this argument of Wittgenstein's as probative.

17. Actually, TMA was first put forth—extremely clearly, I might add—by Plato, in his dialogue *The Parmenides.*

18. Plato himself encourage this misinterpretation of his own doctrine, since he tried to understand the instantiation-relation in terms of relations (e.g., covering, resemblance, part-hood) that spatiotemporal entities bear to one another. See his dialogue, the *Parmenides.*

19. Some properties would appear to be self-instances (e.g., the property of being a property). But Russell (1908) disagrees.

20. Readers familiar with Chomsky (1965) will see that I'm using his distinction between performance and competence. The argument being put forth is a highly Chomskyan one.

21. "WA" stands for "Wittgenstein's argument."

22. Ayer (1968) makes the same point. Ayer (1968) is, in general, a powerful critique of the private language argument.

23. We use the expression "the English language" to refer to what are in fact distinct, but connected, languages. The language that Chaucer spoke is an ancestor of the language that we speak. The same is true, though less obviously, of the language that Shakespeare spoke. A language, in the strict sense, is a set of semantic rules.

[106]. See Russell (1905, 1917, 1918, 1920, 1948, 1956).

[107]. "DD" is short for "definite descriptions."

[108]. Kripke (1980) puts forth this argument.

[109]. See his book *Psychology and Object-theory.*

[110]. "FR" is short for "Frege."

[111]. Frege (1892).

[112]. "IB" is short for "the inventor of bifocals."

[113]. He used the words "ungerade" and "gerade," which are literally translated as "direct" and "indirect."

9. "IFT" stands for "idea embodied in Frege's theory."

10. See Russell (1905), Frege (1879, 1954), and Taylor (1998). The term "syntactic ambiguity" is due to Russell (1905). The concept of syntactic ambiguity was first clearly analyzed by Frege (1879).

11. An open-sentence containing two or more free variables (e.g., "x loves y," is satisfied by an ordered n-tuple of expressions that satisfy). Thus, the ordered pair <"Romeo," "Juliet"> satisfies "x loves y."

12. In the 1980's, there was a *Saturday Night Live* skit that was based on this very ambiguity.

13. Sometimes generating the narrow-scope reading is a sensitive task, as this example shows. It can't always be done by permuting expressions that even resemble those found at the level of surface-structure.

14. See Quine (1953, 1960, 1966, 1981).

15. "QA" stands for "Quine's argument."

16. This is the principle that if x is the same thing as y, then x has any property that y has (and *vice versa*).

17. Frege was the first to identify the logical forms of quantified statements and to give a clear analysis of the phenomenon of syntactic ambiguity.

[114]. Frege (1879, 1884, 1892b). See also Russell (1920), Quine (1940), Reichenbach (1947), Carnap (1958), and Stoll (1963).

[115]. "SI" is short for "self-identical."

[116]. Frege hoped to formalize ***al*** inferences using quantifiers. He also believed, for reasons that aren't important, that all formal, quantified inferences can be represented as inferences whose constituent sentences use only two quantifiers, these being "all" and "some" (or, respectively, "for all x" and "for some x").

He got far. But it turned out some quantified inferences cannot be represented as inferences involving no quantifiers besides the two just mentioned. For example, Dana Scott proved that inferences involving "most" can't be formalized in this way.

[117]. "RP" stands for "reparsed."

[118]. "S" stands for "symbolized."

[119]. The "G" stands for "generalized."

[120]. The "AA" stands for "alternate argument."

[121]. Frege (1954).

[122]. Part of why he believed this, I think, is that he falsely identified formal truth with analytic truth, and thus thought that, if arithmetical truths weren't formally true, it would follow that they were synthetic. Given his (totally correctly) held that arithmetical truths are not synthetic, he was forced to hold (falsely, as we'll see) that arithmetical truths are formally true.

[123]. Immanuel Kant (1724–1804) held that all non-formal truths are synthetic. And since he recognized that

"2 + 2 = 4" and other arithmetical statements are informal truths, he held that they were synthetic (i.e., that they didn't hold entirely in virtue of the structures of the concepts involved in them.) He said that

"7 + 5 = 12" is a *synthetic* truth (i.e., that it doesn't hold in virtue of its meaning).

This theory of Kant's was a disaster for philosophy. "7 + 5 = 12" is not an *empirical* truth. It has nothing to do with physical operations. That's why, if putting a seven-pound weight on top of a five-pound weight that's on an otherwise unoccupied balancing pan yields a total weight of *more* than twelve pounds, we don't say that $7 + 5 \neq 12$. Instead we say that weight isn't a strictly

additive quantity. And that's precisely what physicists *do* say, since it is a consequence of Relativity Theory that, indeed, the total poundage of those two weights is just slightly *over* twelve lbs.

Kant saw that "7 + 5 = 12" is not empirical. He should have concluded that it is analytic. But Kant had a very narrow, and also a completely incoherent, conception of analyticity. He thought that all analytic truths are formal truths. (Why this is false and incoherent is discussed in Chapter 18.) Kant was thus stuck with the problem of explaining how we could have non-empirical knowledge of non-analytic truths. (Kant used the term "a priori," instead of "non-empirical.")

But this problem is unsolvable. Analytic truths hold entirely in virtue of their meanings. Synthetic truths don't and therefore hold in virtue of meaning-independent facts. So a synthetic, non-empirical truth is one that holds neither in virtue of its meaning nor in virtue of virtue of meaning-independent facts and therefore isn't a truth at all.

Kant saw this. But instead of drawing the correct inference—namely, that there is no synthetic a priori truth—he draws the conclusion that "7 + 5 = 12" is a fact about our minds that we project into reality. And, true to this position, Kant says that, for creatures with different cognitive architectures, "7 + 5 = 12" isn't true. So, for Kant, "7 + 5 = 12" isn't really true at all.

Kant's position is desperate and implausible. Frege put a swift and fitting end to Kant's chicanery. Frege saw that, even though it is not a formal truth, "7 + 5 = 12" is analytically true. Frege thus discovered the profoundly important fact that there are analytic, non-trivial truths.

With this discovery analytic philosophy began. Moritz Schlick (1882–1936), an early analytic philosopher of extreme ability, if not genius, said that the essence of analytic philosophy could be distilled into one slogan: "no synthetic a priori." See Schlick (1934). Philosophy *is* the study of informal analytic truth. Formal truths fall within the jurisdiction of mathematics. Synthetic truths fall within the jurisdiction of empirical science. Informal non-empirical truths fall within the jurisdiction of philosophy. Are there such truths? Yes. For any given formal truth (e.g., ‹ if P and Q, then Q›), there are infinitely many informal analytic truths.

[124]. Russell (1920).

[125]. Frege (1954).

[126]. (2) contains the quantifiers "something x" and "something y." Since "something x is an apple on the table" is equivalent with "at least one thing is an apple on the table" it might seem as though (2) *does* (implicitly) contain an expression referring to the number one. But this isn't so, since "something x is an apple on the table" can be rewritten as *A and B are not mutually exclusive properties,* where A(B) is the properties possession of which is necessary and sufficient for being an apple (being on top of the table).

[127]. Frege (1954).

[128]. I'm modifying Frege's solution to avoid irrelevant technicalities.

[129]. This terminology is borrowed from Russell (1920), who borrowed it from Dedekind (1888).

[130]. Bertrand Russell (1903) was the first to make this clear. Russell's insight was a refinement of one due to Richard Dedekind (1831–1916).

[131]. In this section, the word "logic" should, whenever the context requires it, be taken to mean "logic *plus* set theory." See note 20 for clarification.

[132]. A sentence is ***formally true*** iff every sentence having the same form is true. "Either Smith is tall or it is not the case that Smith is tall" (henceforth "ST") is formally true because every sentence of the form ‹P or not-P› is true. "Any case of knowledge is a case justified belief" (henceforth "KJ") is *not* formally true because it has the same form as "any case of ignorance is a case of justified belief" which is false. KJ is ***analytically true*** (or ***analytic***) because, setting aside the empirical knowledge needed to grasp it, no empirical knowledge is needed to know that it is true. That it is true is to be seen by understanding the relevant facts about the concepts composing its meaning; empirical inquiry is irrelevant.

In addition to being formally true, ST is analytically true. Setting aside the empirical knowledge needed to grasp ST itself, no empirical knowledge is needed to know whether it is true or false. All formal truths are analytic truths; but not all analytic truths are formal truths (e.g., KJ is analytic but not formal).

[133]. During its hey-day, formal logic wasn't clearly distinguished from set-theory; it was thought that the latter is a branch of the former. The great pioneer logicists—Frege, Russell, Whitehead—attempted to reduce arithmetic to set theory, which they wrongly described as "logic."

[134]. For reasons given in Chapter 4, I believe that it's sentence-tokens that are formally true, the reason being that it's ***sentence-tokens,*** not sentences *per se* (i.e., not ***sentence-types***), that are either true or false to begin with. (In what remains of this chapter, including this footnote, take "sentence" to mean "sentence-type" *or* "sentence-token," depending on how the context in question suggests that it ought to be disambiguated.) I also believe that, although formal truth is a property of sentences, there is a non-sentential analogue of it; there is a property that propositions (sentence-meanings, as opposed to sentences) have that is similar to the property of being formally true.

[135]. "TPT" is short for "two plus two."

[136]. I repeat my previous point that, in this context, "sentence" is ambiguous between "sentence-type" or "sentence-token" and is to be disambiguated charitably.

[137]. He did this in his now famous (1931) paper "On Formally Undecidable Propositions of *Principia Mathematica* and Related Systems," which is available online. But I recommend reading a summary of it beforehand. In his book *The Infinite,* A.W. Moore clearly and concisely states the main ideas. The outline about to be presented of Gödel's argument borrows heavily from one of the two outlines that Moore presents of it in that book. Gödel's argument was, to some extent, modeled on George Cantor's brilliant proof that, although there are infinitely whole numbers and also infinitely many real numbers, there are *more* real numbers than there are whole numbers. Moore makes it more clear than any other author I've read, including Gödel himself, how exactly Gödel's proof is modeled on Cantor's. And the outline about to be given owes much to Moore's crisp and lucid discussion of it.

A point made by Richard Dedekind will help us outline Cantor's striking insights into the concept of an infinitely large set. S is an infinitely large set iff it can be put into a one-one (read: "one to one") correspondence with a proper subset of itself. Consider the set of whole numbers. The members of that set can be put into a one-one correspondence with the set of even numbers. (Assign 1 to 2, 2 to 4, etc.) Given that there are infinitely many even numbers, it follows that the set of whole numbers is infinitely large.

Now for Cantor's insights. The set of whole numbers is equinumerous with (has the same

number of members as) the set of rationals. Let L be the following list: 1/1, 1/2, 2/2, 1/3, 2/3, 3/3, 1/4, 2/4, 3/4, 4/4, etc. Given any rational number R, R will occur somewhere on L. It's obvious that L's members can put into a one-one correspondence with the members of the set of whole numbers. (Assign 1 to 1/1, 2 to 1/2, etc.) Therefore, counterintuitive though it may seem, there are exactly as many whole numbers as there are rationals.

But there are more real numbers than there are whole numbers. One set S is larger than another set S* iff, supposing that each member of S* is paired off with exactly one member of S, some member of S will remain unpaired with any member of S*. Bearing this in mind, suppose for argument's sake that LR is a list of all the real numbers. Each real may be represented as an infinitely long series of decimals. (So the real number corresponding to the whole number 5 may be represented as "5.0000000 ‹repeating›.") Let F be a function that is defined as follows. For each number n, if the nth digit on the nth item on LR is a "5," F replaces the "5" with a "4," and if that item is *not* a "5", F replaces it with a "5." This procedure defines an infinitely long series of digits and thus corresponds to some real number RN. But the first decimal of RN differs from the 1st digit composing the first entry on LR, and the second decimal of RN differs from the 2nd digit composing the second entry on LR, and so on. It follows that there is no entry on LR such that RN is identical with that entry. If we add RN to LR, we can simply repeat the procedure just described, and the result will be *another* real that isn't on the list. Thus, any attempt to "enumerate" (make a list of) the reals falls short: there is always some real that can't be tagged. Alternatively, the supposition that a given list LR is the list of all the reals entails that there is some number RN that *isn't* on that list; and, consequently, that supposition is incoherent. So there are more reals than there are whole numbers. Since, as we saw, there are as many wholes as there are rationals, it follows that there are more reals than there are rationals.

[138]. George Berkeley foreshadowed the use of this technique when he said that for x and y to be instances of some property P is for x to resemble y in some respect. See Chapter 2. For a clear discussion of the concept of definition by abstraction, see Reichenbach (1947, p. 248).

[139]. See Section 2.3.

[140]. See Russell (1948).

[141]. Many of these qualifications relate to a stunning discovery made by Alonzo Church (1901–1995) and, working independently of Church, Alan Turing. Church and Turing showed that there is no decision procedure for ***first order logic.*** (See Quine (1995) for an excellent discussion.) First order logic is the branch of logic that studies those dependence-relations holding among sentences that can be understood in terms of the concepts expressed by the following expressions: "or," "and," "not," "for some," and "for all."

[142]. Including Jerry Fodor (1975, 1987) and Zenon Pylyshin (1984). I regard this view as deeply misguided. See Kuczynski (2002, 2004, 2005, 2006, 2006b, and 2007).

[143]. But not sufficient. See Kuczynski (2006b).

[144]. See Gödel (1953).

[145]. It should be pointed out that relations can be thought of as properties—as properties of ordered-pairs (or, more generally, n-tuples) of objects. The relation of the being taller than can be thought of as a property had by the ordered pair <x, y> exactly if x is taller than y. The relation of being in

between can be thought of as a property had by the ordered triple <x, y, z> exactly if x is in between y and z. And so on. Instances of relations *are* instances of properties. There is therefore nothing to seeing that an instance of one property (e.g., that of being green) bears a certain relation (e.g., that of being on top of) to an instance of some other property (e.g., that of being orange) other than your seeing there to an instance of some third property. It follows that your seeing Pat is identical with your seeing various property-instances and is therefore identical with your being visually given an existence-claim (to the effect that the corresponding properties are instantiated).

[146]. "F1" is short for "Fred 1."

[147]. See Mill: *A System of Logic.*

[148]. "BT" is short for "bolo tie."

[149]. Kripke (1980) deserves credit for introducing this vitally important distinction to philosophy.

[150]. Frege (1892).

[151]. Russell (1905, 1917, 1948).

[152]. See Kripke (1980). Kripke was the first to make many of these important points.

[153]. Salmon (2007).

[154]. The argument just outlined is thoroughly discussed in Chapter 8. Actually, "Socrates" describes Socrates even less than your social security number describes you. That number encodes information about when you were borne (etc.), whereas "Socrates" doesn't even do that.

[155]. We're setting aside the irrelevant fact that there probably is somebody *other* than the great statesman in questing who is named "Winston Churchill." The problem we are dealing with has to do with the fact that, when it's taken to refer to the wartime British prime minister, "Winston Churchill" does not, at this point in time, have an existing referent.

[156]. See Kuczynski (2007) for an extensive discussion of the nature of this property of being identical with JMK (or any other specific individual). But a word on it may be appropriate here. My existence is not a brute fact. Various events are constitutively involved. These events are of various

kind, ranging in nature from the sub-atomic to the psychological. These events are organized; they collectively form a structure. I will continue to exist as long as that swarm of events retains that structure, and no longer. The swarm of events that currently constitute me is an installment in a series of such swarms whose origins ultimately lie in the union of a certain sperm and egg. In light of these points, let P be the property of being a swarm of events of the just-described kind having the just-described origin. In that case, my existing is indistinguishable from P's being instantiated.

[157]. It might seem that PTP and PJMK are redundant in this context. P_1, it is readily seen, is true iff PJMT is instantiated. (If PJMT is instantiated, PTP and PJMK are automatically instantiated.) So why need we identify P_1 with P#? Why not identify it with a set containing just PJMK? Propositions must have certain structural properties in common with the sentences that express them. (The reasons for this are given in Chapters 3 and 4.) P# has the relevant structural similarities with S_1. (Notice that there is a one-one correspondence between P#'s membership and S_1's decomposition.) By, by itself, PJMT doesn't have those properties, and neither does a set containing only PJMT. Because if we identify P_1 with P# given that (for reasons discussed in Chapters 3 and 4) propositions are isomorphic with the sentence.

[158]. "TCS_1" is short for "the truth conditions of S_1."

[159]. "SG" stands for "the story of Gigantus."

[160]. Chapter 3 of Salmon (2005).

[161]. "MO" is short for "mythical object." Salmon's argument is told, not in terms of "Zeus," but in terms of "Maggoty Meg," a fictitious witch. I've made other, equally inconsequential changes.

[162]. "KA" stands for "Kripke's argument."

[163]. "SE" is short for "Smith Earth."

[164]. "TS" is short for "Twin-Smith."

[165]. What a speaker means often diverges from what his words mean. If I tell an employee, "I think it might be a good idea for you to clear your desk," I'm really telling him that he's been fired. But my words don't literally

mean: you're fired. Their literal meaning is innocuous. See Chapter 4 for a discussion of this.

[166]. See Burge (1979, 1982).

[167]. Putnam (1996).

[168]. "BA" is short for "Burge's anti-individualism." In a moment, I'll define the term "anti-individualism."

[169]. "AU" is short for "arthritis unpleasant."

[170]. See Burge (1979, 1982). Evans (1982) puts forth a similar argument.

[171]. "U_1" is short for "utterance #1."

[172]. "MP" is short for "Max plaid."

[173]. "VR" is short for "visual perception of robot."

[174]. "CVR" is short for "content of" "VR."

[175]. See also Russell (1948: 271–294).

[176]. Gareth Evans (1982) was the first, as I know, to see that Burge's position has this consequence. But Evans, being a content-externalist, accepts this consequence, deeply bizarre though it is.

[177]. "MP" is short for "Max plaid."

[178]. Strawson (1950) makes a similar point.

[179]. "BV" is short for "Burge's view."

[180]. Wittgenstein (1958) repeatedly says that thought isn't a psychological (or, as he puts it, an "inward") process. Wittgenstein's argument's for this ludicrous position are thoroughly examined in Chapter 5. Burge's argumentation, though deeply flawed in my view, is considerably better than Wittgenstein's.

[181]. See Burge (1986).

[182]. See Falvey and Owns (1994) for an example of the lengths to which content-externalists have gone to validate their clearly degenerating research-program.

[183]. See Kuczynski (2007) for a thorough discussion.

[184]. See, for example, Bertrand Russell's 1948 book *Human Knowledge: Its Scope and Limits*.

[185]. See Gettier (1963). These very examples are Gettier's.

[186]. See Chapter 7.

[187]. I am using the word "laws" broadly in this context.

[188]. See Lewis (1973, 1984) and Kment (2007).

[189]. See Churchland (1988).

[190]. Here I am heavily modifying and developing the idea that Ronald Dworkin (1977) made, and put to excellent use, in a totally different context.

[191]. The whole concept of defeasibility is alien to empiricism, since it involves the concepts of power, which is one that cannot be understood in empiricist terms.

[192]. This argument was put forth by John Stuart Mill in *A System of Logic* .

[193]. See, for example, Searle (1992).

[194]. See Freud (1915) for a telescoped rendition of a similar line of thought.

[195]. Russell (1917) puts forth a similar argument. Tyler Burge rejects the position for which I am arguing. See Burge (2007).

[196]. "KH" is short for "knowledge how."

[197]. Karl Popper (1902–1994) says this in *The Logic of Scientific Discovery.* Hans Reichenbach (1891–1953) says it in *Experience and Prediction.* Rudolf Carnap (1891–1970) says it is in *The Philosophy of Science.* And Carl Hempel (1905–1997) says it in *The Philosophy of Natural Science* and in *Aspects of Scientific Explanation.*

[198]. By "prove" he meant "justify."

[199]. Empiricism is rooted in a failure to distinguish knowledge-conducive thought from the things that trigger it. (This is shown in Chapter 13.)

[200]. I can only speculate as to why this is so. But it is, I suspect, because philosophers of science tend to be extremely dismissive of any viewpoint that grants the existence of unconscious thought or otherwise makes concessions to psychoanalytic thought.

[201]. Hilbert (1999) makes it clear why Euclid's efforts don't quite succeed.

[202] .Moritz Schlick (1934), Russell (1984), and Roderick Chisholm (1976)

were foundationalists.

[203]. Otto Neurath (1932) was a coherentist. So is Laurence Bonjour (1985).

[204]. Advocates include: Plato, Descartes, Leibniz, Karl Popper, and Saul Kripke.

[205]. Advocates include: Aristotle, Thomas Hobbes, J.J. Smart, Hilary Putnam, and Jerry Fodor.

[206]. Spinoza advocated this view.

[207]. Quine (1960) advocates this view.

[208]. In *The Principles of Human Knowledge,* George Berkeley. Other advocates include Bertrand Russell (*The Analysis of Matter*) and William James (*Essays in Radical Empiricism*).

[209]. In the *Monadology,* Leibniz gives a similar argument, and so does Gottlob Frege (1918).

[210]. Frank Jackson forcefully makes this point in several articles. See Jackson (1998).

[211]. In his paper, "Mental Events," Donald Davidson argues for this position, or one close to. We'll consider Davidson's argument in Section 9.2.1.

[212]. Joseph Levine has published many papers in which he argues that there is an explanatory gap. See his paper "Conceivability, Identity, and the Explanatory Gap," which is available online at: *http://cognet.mit.edu/posters/TUCSON3/Levine.html*

[213]. See Hempel (1966).

[214]. In his book *Essays on Actions and Events.*

[215]. "DA" is short for Davidson's argument.

[216]. The following argument is due to Quine (p. 153 of *From a Logical Point of View*).

[217]. Hempel (1952) and Pap (1958) made this point. Kripke (1972) and Putnam (1975) put it to good use.

[218]. J.J. Smart (1959) gives this argument. So does Fodor (1968).

[219]. C.D. Broad makes a similar point in *The Mind and Its Place in Nature.*

[220]. My own view is that this reasoning is solid. I say this despite my unshakable, but—so far as I know now— very hard to defend suspicion

that dualism cannot be completely wrong.

[221]. In Chapter 17, it is said what exactly it means to identify objects with causal sequences, and it is also said why exactly objects must be identified such sequences.

[222]. Fodor (1968).

[223]. Richard Rorty (1979) makes this point.

[224]. These two meanings are clearly distinguished by John Searle in *The Rediscovery of the Mind.*

[225]. See Watson (1924) and Carnap (1934).

[226]. "RS" is short for "Ramsey sentence." RS is given by David Lewis (1972). In his paper "Psychophysical and Theoretical Identifications." David Lewis (1983). See also Frank Ramsey's paper "Theories," in Ramsey (1990) and "The Ramsey Sentence," Carnap (1966).

[227]. "HA" stands for "Hume's argument."

[228]. See Hume *A Treatise Concerning Human Nature* Book 1, Part 4.

[229]. Reichenbach (1956) makes a similar point.

[230]. "HR" is short for "Hume's response."

[231]. See Infeld (1950).

[232]. See Popper (1983).

[233]. At least so far as their various beliefs form a consistent set—we'll set aside the secondary question of how, given discrepancies among their views, the right ones are to be selected.

[234]. Speaking frankly, I'm not 100% that I believe what I'm saying in this passage. I believe that it's a defensible position. But I *do* believe there to be a connection between an explanation's being good and it's being likely to be correct. I simply haven't been able to identify that connection. As a result, the attempt to refute skepticism put forth in this chapter is not adequate. But I have left this chapter in, because I think that components of it are cogent and because, were I to get rid of the non-cogent parts, the merits of the cogent parts would be unclear. (I believe, for example, that my refutation of Hume's analysis of inductive inference is cogent; and I believe the same to be true of my analysis of scientific explanation. What I don't believe to be cogent is my attempt to base a refutation of skepticism

on that attempt.)

[235]. See Quine (1953, 1960, 1966, 1981).

[236]. Whether this point is consistent with the points I make in Chapter 25 is a delicate issue.

[237]. In point of fact, the laws of physics *do* demand that the collective weight of two rocks, each weighing exactly 1 pound, is just slightly more than 2 pounds. If you mix a quart of water with a temperature of 70° with a quart of water with a temperature of 90°, the result is not a body of water with a temperature of 160°. So temperature is not an "additive" property. If the object formed by conjoining two 1-pound rocks weighed more than 2 pounds, then weight would not be an additive property. Since Relativity Theory shows that two conjoined 1-pound rocks weigh slightly more than 2 pounds, weight is *not* a strictly additive property.

[238]. There are some trivial exceptions to this. For example, given that, at past time t, Eddy had two cars, one can infer that, in the future, any given object will be such that, at t, Eddy had two cars. The empiricist holds that, given any proposition P concerning the past, there is no proposition that isn't analytically entailed by P such that Q concerns the future and such that, supposing P to be true, an acceptance of Q is *ipso facto* justified.

[239]. Berkeley makes this point in his (1734) book *The Principle s of Human Knowledge.* Wittgenstein (1958) makes some remarks that could be interpreted as attempts to make this point.

[240]. His arguments for this are put forth in his book *Three Dialogues between Hylas and Philonous.* Some of those arguments are elucidated in his book *the Principles of Human Knowledge.*

[241]. For example, Anthony Brueckner, a contemporary philosopher, believes that skepticism is irrefutable, and he has spent much of his career trying to undermine attempts to refute it. He does this on a case by case basis. In other words, he considers individual attempts to refute skepticism, and tries to undermine them. To my knowledge (which, in this domain, is limited), he has not provided a general argument as to why skepticism is irrefutable. My personal feeling is that we would learn more from such an argument—even a failed one—than we would from the alleged failure of specific attempts to refute skepticism.

[242]. The works of Berkeley's relevant to this section are his *Principle s of Human Knowledge* and his *Dialogues between Hylas and Philonous.*

[243]. The argument just given is reconstructive. Berkeley himself doesn't give it. But it provides the justification for assumptions that Berkeley makes without fully justifying.

[244]. Here's an argument for this thesis. (It isn't an argument Berkeley gives.) Consider my two house-perceptions. (Let PB and PS, respectively, be the perception that tells me that the house is big and the perception that tells me that the house is small.) According to the procedure just described, I'd have to throw out one of those two perceptions. And, given any pair of conflicting perceptions, I'd have to do the same thing. In no case would there be any principled way of doing this. Which means that, even if I happened to be right in my choice—i.e., even if I happened to select the perception that was in fact wrong—I'd be doing so haphazardly, and therefore wouldn't *know* that I was making the right choice. Which means that we wouldn't know that our perceptions (or the ones we chose, rather) were correct, even if they were. Which means that those perceptions wouldn't give us knowledge.

[245]. This is the argument Berkeley puts forth, though not as speedily, in his *Dialogues between Hylas and Philonous.*

[246]. Technically, this is an inaccurate way of putting it because it suggests, falsely, that there are absolute rates of change. But the substance of this point is correct.

[247]. He was not the first to do this. Galileo (1564–1642) did it before him, and so did Lucretius (99 B.C.–55 B.C.).

[248]. Strictly speaking, shape *per se* isn't a primary property. What is a primary property is some *specific shape.*Similarly, taste *per se* isn't a secondary property. What *is* a secondary property is some specific taste—e.g., the taste of a Snicker's bar. The same thing *mutatis mutandis* holds for the other items on these two lists.

[249]. Judging by certain passages in his work, Berkeley himself seems to make this claim.

[250]. Actually, Hume would say otherwise; and I'm not entirely sure that he's

wrong. (See Chapter 16.) But we can set this aside here.

[251]. In his *Essay Concerning Human Understanding,* Locke made essentially this point.

[252]. If rotten meat had had S*, instead of S, people wouldn't have minded keeping it in their caves (or whatnot), and would have been more willing to eat it. Food often smells rotten before it tastes rotten. But this (alleged) fact is irrelevant in this context, since we're talking about the epistemological, not the emotional, virtues of odors.

[253]. SA

[254]. Since you're much smarter than those animals, you can *do* a lot more with the relative of paucity of information given to you by your low-res perception than those animals could do with their high-res perceptions. But that has nothing to do with what those perceptions *themselves* are saying.

[255]. Because the particles composing very hot things move so fast, those particles do a lot of damage: they damage nerve-endings and other organic tissues. Hence the unpleasantness associated with touching such things.

[256]. "BR" is short for "Berkeley's reasoning."

[257]. Actually, Berkeley claims—totally implausibly—that people agree with his contention that there are no trans-perceptual objects.

[258]. "Defeasible" means "capable of being overridden." Defeasible evidence is evidence that is inconclusive but is to be given credence until outweighed by contrary evidence. Most evidence is defeasible. The only possible cases of non-defeasible evidence are things that are evidence of themselves, e.g., headaches, tickles.

But, since the idea of a thing's being evidence of itself is of doubtful coherence, it's likely that all evidence is defeasible.

[259]. Rudolf Carnap (1890–1971) spent years trying to translate perception-statements into object-statements. His efforts culminated in his (1928) book *The Logical Structure of the World.* At no point in that book, or in any other work of his, does Carnap acknowledge the fact that one's

perceptions are evidence of the existence of things outside of one's own mind *only* because those perceptions tell one that there exist such things. Carnap made many attempts to vindicate phenomenalism (the view that statements about physical objects could be translated, without loss of meaning, into statements about one's own perceptions and sensations). Every attempt he made was subject to obvious counterexamples. When these were pointed out to him, he chose *not* to jettison phenomenalism, but to tinker with the technical details of his particular attempts to establish the truth of it.

Carnap was a hardcore empiricism—hence his tenacious advocacy of phenomenalism. He felt that empiricism was the "scientific" view to take —that not being an empiricist meant being "metaphysical." (He took it for granted that being "metaphysical" is bad. "Why is it bad?", one might ask. "Because," says Carnap "it's unempirical to be metaphysical." This vicious circularity is the most Carnap does in the way of clarifying his view that "metaphysics" is bad and of defending of his view that empiricism is good.) At a conference in Mexico in 1968, Carnap declared that, contrary to what he had been assuming for his entire career, there is no such thing as science. He advocated the post-modern view that science is whatever works for one, given one's personal objectives.

Carnap is often praised for his willingness to change his mind in the face of new information that didn't bear out his theories; and his bold (1968) rejection of science is seen as epitomizing his (alleged) openness of thought. But it seems to me that, in doing this (pseudo-)180°, Carnap is *not* changing his mind. He's doing the exact opposite. He's saying that, since science doesn't fit his rather idiosyncratic theory as to what science is, science doesn't exist. It should be pointed out that John Stuart Mill (1806–1873) had advocated the very conception of science advocated by Carnap (minus the distinctively 20th century forays into formal logic) and, in addition, that William Whewell (1794–1866) had put forth cogent arguments against Mill's view.

[260]. To my knowledge (which is limited, given that I'm not a Berkeley-scholar), Berkeley is never given credit for making these discoveries. I am, so far as I know, the first author to do so. But there is no doubt that he

made them. See his work *The Principles of Human Knowledge.*

[261]. This is a point that Berkeley himself makes very clearly on several occasions. See his *Dialogues Between Hylas and Philonous*.

[262]. In this context, I'm not using the expression 'deep structure' in the Chomskyan sense.

[263]. A sentence is 'perspicuous' if it is clear *n i the sense that* its inferential structure (what it entails and what entails it) can be read off of its grammatical structure. See the entry for 'perspicuous' in the second appendix.

[264]. Another deeply insightful and prophetic point that Berkeley makes (in the *Principles of Human Knowledge*), and for which he provides reasonably cogent argumentation, is that there is no empirical basis for the claim that position in either space or in time is absolute. He says, rightly, the spatiotemporal position must be understood in comparative terms. (One event e is later than event e*; one object x is further from y than is z.) But to his discredit, Berkeley doesn't address Newton's famous argument to the contrary. For argument's sake, suppose that motion (and therefore position) is relative. In that case, there is no difference between your standing still while some nearby bucket spins around and the bucket's spinning around while you stand still. But, says Newton, there is a difference. If you stand still and the bucket spins around, the surface of the water in the bucket becomes concave. But it doesn't become concave if the bucket remains still while you run around it.

What does this show? First of all, we must distinguish between *kinematic* and *dynamic* relativity. ("Kinematic" means "having to do with how things are moving." "Dynamic" means "having to do with the nature and intensity of the forces that are at work.") Kinematically, your relationship to the bucket (when it's stationary and you run around it) is just like the bucket's relationship to you (when it is spinning and you are stationary). But *dynamically* the two situations are not symmetrical. And, Newton very reasonably inferred, the fact that these two situations are dynamically non-comparable shows that, when the water in the bucket becomes concave, the bucket really is moving. In other words, it is

moving relative to the coordinate system with respect to which given any event, that event would be assigned a position in time and space even if it were the only event that ever occurred.

In general, the effects of *accele rated motion* (speeding up, slowing down, changing direction) tell us what is really moving and what is moving in, at most, a relative sense.

To take another example: Smith is standing by the side of the road. Jones is driving on that road, and he is accelerating very quickly. Kinematically, Smith's relationship to Jones is just like Jones' relationship to Smith. But dynamically the one person's situation is very different from the other's. Jones will feel the g-forces. Smith will not. This difference, Newton would say, shows us that Jones really is moving.

Newton's reasoning is extremely formidable. Berkeley didn't address Newton's argument. (His responses are so curt that he either seems not to have understood what Newton was saying or he deliberately turned a blind eye to it, since he knew he couldn't parry Newton. In any case, Berkeley was vindicated—after a fashion—when, working two hundred years after Newton, Einstein showed that motion is not only kinematically, but also dynamically, relative.

[265]. Karl Popper had a lot to say about the limits of predictability. What I'm saying here doesn't always coincide with what he says, but much of it is similar to it. See Popper (1983, 1992).

[266]. No time would have elapsed unless there had been a change. Time is a relation between events. No changes, no events. No events, no time. The fact that time is to be understood in this way spells doom for Hume's famous eliminitavist analysis of causation and for a number of doctrines in epistemology and the philosophy of science to which it has given rise. See Chapter 17.

[267]. We're setting aside trivial correlations; e.g., "anything that is identical with the headache I'm now feeling is also identical with something I want to get rid of."

[268]. See Freud (1949).

[269]. The straightforward version ("you are free if you can do what you want to do, and unfree so far as you cannot") was advocated by John Locke (1632–1704) and G.E. Moore (1873–1958). The more sophisticated version is advocated by Immanuel Kant (1724–1804) (in the *Critique of Practical Reason*), Sigmund Freud (1856–1941) (in *An Outline of Psychoanalysis* and *the Question of Lay Analysis*) and Harry Frankfurt (1929–) (in *The Importance of what we care about*).

[270]. Moore (1911).

[271]. In philosophy, split infinitives are hard to avoid, since the consequences of avoiding splitting them are often sentences that don't mean what one wants to say. There's a difference between saying "Smith chose to *not* strike Harry" and "Smith did not choose to strike Harry." The latter says that Smith failed to make a choice; the former says that he did make a choice. Of course, "Smith chose *not* to strike Harry" does avoid the split infinitive. But having the "to" adds oft-needed emphasis.

[272]. See his *Essay Concerning Human Understanding,* Chapter XXI.

[273]. See Frankfurt (1988).

[274]. "SM" stands for "Smith microchip."

[275]. This is thoroughly discussed in Chapter 17.

[276]. "PAP" stands for "principle of alternate possibilities." Frankfurt coined this term.

[277]. There are exceptions to this. For example, counter*logicals* aren't causal claims. An example of a counter-logical would be:

(*) "if that thing were a pentagon, it would have five sides instead."

(*) obviously isn't a causal claim. (*)'s meaning parallels (JFK)'s. Neither says anything about other worlds.

Each merely registers some dependency that holds in this world.

[278]. "FA" is short for "free act."

[279]. See Frankfurt (1988, Chapter 2).

[280]. Frankfurt's analysis of freedom corresponds very closely to the conception of freedom underlying the work of Freud and other thinkers in the psychoanalytic tradition. But Frankfurt was, to my knowledge, the first person to articulate this conception of freedom, and he was the first to defend it in a systematic and rigorous manner.

[281]. "FD" can be thought of as short for "freedom to do what one wants."

[282]. Addiction doesn't obliterate a person's personality; it only temporarily obscures it. The addiction may lead to some character remodeling. But, in most cases, the person is still there, even at the height of the addiction, granting that it may be well-hidden at that point.

[283]. "SMP" is short for "Sartre, Merleau-Ponty."

[284]. See Solomon (460–173).

[285]. Cf. Merleau-Ponty (Solomon 467): "It is inconceivable that we should be free in certain actions and determined in others... [In connection with the will] we ought, therefore, to reject not only the idea of causality, but also that of motivation."

[286]. Freud, Stekel, and other early psychoanalysts referred to OCD as "obsessional neurosis."

[287]. See Davies (2001, 2001b).

[288]. A colleague of mine (who wishes to remain anonymous) told me that sickle cell anemia is adaptive in some respects. My claim that physical ailments aren't adaptive must obviously be qualified. Many ailments involve over-active immune systems and can be beneficial (in certain respects, at certain junctures).

[289] This must be qualified. It's usually not a good idea to look *primarily* to legal principles for answers to philosophical questions. But a knowledge of legal principles can help one fine-tune the answers one has already come up with. For example, many of the laws concerning the admissibility into legal proceedings of certain kinds of evidence do embody real insight into the structure of knowledge.

[290]. In his (1949) work *the Philosophic al Investigations,* Ludwig

Wittgenstein said that all philosophical problems are to be solved by looking at how words are used. (In Chapter 1, it is argued that this view is untenable.) This work of Wittgenstein's initiated a philosophical methodology known as "ordinary language philosophy." (Among the proponents of ordinary language philosophy are J.L. Austin (1911–1960), Norman Malcom (1911–1990), and Peter Strawson (1919–2006).) Ordinary language philosophers—including the ones just mentioned—put a lot of stock in arguments such as the one just given. Their position would be that, since we can refer to corpses by the names of their former occupants, those corpses *are* those occupants. Although there is much to be learned from how use words, it is quite absurd, in my judgment, to hold that philosophical problems are categorically to be solved by studying linguistic behavior.

[291]. In his (1973) book *Problems of the Self,* Bernard Williams (1929–2003) puts forth a similar argument.

[292]. Quine (1960) advocates this view.

[293]. In his (1956) book *Elements of Symbolic Logic,* Rudolf Carnap uses the term "time-slice" (or its German equivalent, rather) in this way. To my limited knowledge, he was the first to do so.

[294]. See Parfit (1984).

[295]. See Chapter 13.

[296]. See Chapter 17 for a discussion of how, exactly, Hume's views on these matters were conservative.

[297]. Hume (1740: 252).

[298]. Hume (1740: 252).

[299]. Defeasible" means "capable of being overridden or proven false." Etymologically, it means "capable of being defeated."

[300]. Also, any given spatiotemporal thing is moving relative to something.

[301]. Obviously (i) isn't *formally* false. It isn't like "snow is white and it is not the case that snow is white." Nor is (i) tautologously false. It isn't like t 'here are not three feet in a yard. (i) is *analytically* false. It embodies an

incoherent assumptions about the nature of persistence in time. For further discussion of the differences between formal and analytic truth, and between analytic and tautologous truth, see Chapters 1, 6, 18, and Appendix 1.

[302]. To use Fodor's apt expression. See his paper "The Present Status of the Innateness Controversy," in Fodor (1981).

[303]. Kant makes a similar point at the very beginning of the *Critique of Pure Reason.*

[304]. Carnap (1966) makes this point. Carnap uses a similar example.

[305]. See Mackie (1980).

[306]. He did this in the *Metaphysics.*

[307]. 'But what about the 'fact' that 1 + 1 = 2. That isn't a distribution of mass-energy?' That is true. But, if the word 'fact' is to be of any philosophical use, it has to be disciplined a bit. And in quasi-stipulatively saying that 'facts' are distributions of mass-energy, I am deviating from ordinary usage only as much as is necessary to allow that word to have role in philosophical discourse.

[308]. In Chapter 11, functionalism is discussed at length, that being why my discussion of it here isn't as rich in illustrations and clarifications as it would otherwise be.

[309]. See Jackson and Pettit (2004).

[310]. "SA" stands for "Smith arson."

[311]. Of course, "smaller" doesn't always equal harder to know. It's easier to know people than it is to know galaxies.

[312]. "HC" is short for "Hume on causality." The "1" corresponds to the fact that this is the first of two analyses of causation that Hume gives.

[313]. Hume, *A Treatise of Human Nature,* Book 1, Chapter XIV.

[314]. "HC" is short for "Hume on causality." The "1" corresponds to the fact that this is the first of two analyses of causation that Hume gives.

[315]. See Infeld (1950), Einstein (1952), Reichenbach (1957), Einstein and Infeld (1967), and Sklar (1974).

[316]. A consequence of the fact that order in time and in space are to be

understood in causal terms is that there are place-times, but no places and no times (i.e., there are places that can be occupied by things existing at different times or times that can be occupied by things existing in different places).

As we noted, one event's preceding another consists in there being a possible signal (e.g., a light beam) beginning with the first and ending with the second. Given a pair of events related to each other in this way, let us say that the one from which the signal is sent can "influence" the one to which it is sent. In light of this, imagine the following. There is some event E_1 from which three different sequences of events

(event-sequences) emanate; these being $E_2...E_1$, $E^*_2...E^*_m$, and $E_{\#1}...E_{\#1}$. E_5 may be able to influence E^*_5 while not being able to influence E#5. In fact, E*5 may be able to influence E*5 but *not* be able to influence E*4. Intuitively speaking, the reason for this is that whether one event can influence another is a function not just of the times of those events, but also of their positions. Thus, if x bears the influence-relation to y, that is in virtue of the relation borne by x's place-time to y's place-time. Thus, order in time is order in place-time, and so, for similar reasons, is order in time.

[317]. See Reichenbach (1957) and Sklar (1974).

[318]. See Reichenbach (1957) and Sklar (1974).

[319]. "IP" stands for "immediately precede."

[320]. See Lewis (1984).

[321]. "KN" is short for "Kangaroos with no tails."

[322]. "'LA" is short for "Lewis's analysis of counterfactuals."

[323]. "LC" is short for "Lewis on causality."

[324]. "BU" is short for "balls in the urn."

[325]. "BB" stands for "black ball."

[326]. "WB" stands for "white ball."

[327]. See Lewis (1986).

[328]. "LA" is short for "Lewis' analysis of induction." See Lewis (1973, 1984, 1986).

[329]. "HA" is short for "Hume's argument."

[330]. Goodman (1983).

[331]. Statements are either true or they aren't. Explanations either goes through or they don't. There are no "scientific" statements, and there are no "unscientific" statements.

To the extent that an utterance lacks a clear meaning, it constitutes a failed attempt to make a statement and therefore *isn't* a statement. And to the extent that an utterance has a clear meaning, it's either true or false, and therefore cannot meaningfully be described as "scientific" or "unscientific." If T is true, science has to accept it; if T is false, science has to reject it. Period.

That said, there are distinctly scientific ways of *generating* statements. Given some statement S, assuming that

(MS) ‹S is a scientific statement›

doesn't say the very same thing as

(MS*) ‹S is a statement›,

MS must say S was generated in a certain way—that it was generated in accordance with procedures whose truth-conduciveness has been definitively established, or some such.

Nonetheless, a number of philosophers have argued that there is some property that all and only "scientific" statements have. For example, in his (1931) book *The Logic of Scientific Discovery,* Karl Popper (1902–1994) argued that a statement is scientific iff it is ***falsifiable.*** To say that a statement S is falsifiable is to say that it could in principle be shown to be false on strictly observational grounds. (The sort of falsificationism that Popper advocated is not to be confused with the very different doctrine, discussed in Chapter 1, according to which a statement is *meaningful* iff it is falsifiable.)

To be sure, *some* theories can be falsified on strictly observational grounds. Suppose that, according to theory T, anyone who takes

medication M will be acne-free within five hours. Given that Smith took M at noon, it follows that T is false if Smith still has acne at 5:00 P.M. Given that Smith *does* still have acne at 5:00 P.M.; and given that this fact is painfully obvious to anyone who isn't blind, it follows that T can be falsified on strictly observational grounds.

But T is not representative of most theories in this respect. Whether or not a given theory has a given observable consequence is likely to depend on whether each of various other theories is correct. (To confirm a given astronomical theory, one must assume the truth of various theories concerning the behavior of light. In making this assumption, one is assuming that the instruments used to authenticate those theories were reliable ones. In making *this* assumption, one is, for reasons similar to those just given, assuming the truth of physical theories additional to those already mentioned. And so on and so forth.) So given only that some observable event hasn't occurred, it's not likely to follow that this or that specific theory must be false.

Aware of the fact that theories can't typically be falsified on strictly observational grounds, W.V.O. Quine (1908–2001) took the position that *any* theory is consistent with *any* body of observational data.
Given any theory and any set of observations, there is, says Quine, a way of modeling those observations that is consistent with that theory. Quine refers to this thesis of his as the "Duhem-Quine thesis." (Pierre Duhem (1861–1916) was a physicist and philosopher who Quine wrongly thought to have advocated a similar thesis. In actuality, Duhem advocated a much more circumscribed position.)

The so-called Duhem-Quine thesis is an extremely facile one. As before, suppose that, according to theory T, anyone who takes medication M will be acne-free within five hours. Given that Smith's acne is going strong six hours after taking M, there's nothing we can say that isn't brazenly ad hoc and that doesn't create vast numbers of otherwise non-existent anomalies.

It's obvious and trivial that, given any theory T that concerns the external world, no matter what observations we've had, it's *logically* possible that T is false. So unless it's completely hollow, the Duhem-

Quine thesis can't just say that, for any theory T that concerns the external world, no body of observations analytically entails that T is false. It must therefore be to the effect that, given any body of observations and given any theory T, the explanatory costs of accepting T are no greater and no smaller than the explanatory costs of rejecting T.

But *that* statement is very obviously false. The data being what it is, the explanatory costs of holding onto T far outweigh the explanatory gains of doing so. If we hold onto it, we're stuck with various anomalies that we wouldn't otherwise be stuck with. If we reject it, we're not stuck with those anomalies; nor are we stuck with any other anomalies that we weren't already stuck with. So depending on how it's interpreted, the Duhem-Quine thesis is either trivial or false.

[332]. Hempel (1965).

[333]. "YP" is short for "yellow pencil."

[334]. "MC" stands for "the minimization of causal anomalies."

[335]. "ED" stands for "elimination of discontinuities."

[336]. "OH" is short for "objector's hypothesis."

[337]. In this context, *anything* described by a true sentence counts as a "state of affairs." So facts are "states of affairs," as we're using this expression, and so are occurrences.

[338]. Hempel (1965).

[339]. The "A" stands for "accidental."

[340]. Ayer (1972).

[341]. Hempel (1966).

[342]. The "N" stands for "nomic."

[343]. The "W" stands for "wide-scope," the idea being that the nomic operator is given maximally wide-scope, since the whole of "for any x, if x is metal, then expands" is governed by it.

[344]. "DN" is short for "degrees of necessity."

[345]. Obviously the *sentences* used to express it aren't necessarily true. It's the content of those sentences that cannot fail to be true. The linguistic vehicles can easily fail to be true; even in our universe, they're only

regionally true.

[346]. "HIP" is short for "Hesperus identical with Phosphorous."

[347]. An explanation of this is given in Chapter 8.

[348]. This is only approximately correct. See Appendix 1, Section 4.5.

[349]. The argument about to be given is found in Bonjour (1998) and also in Dummett (1973).

[350]. This is a shame. Quine was a brilliant man. He had a lot to say, and he would have had a lot more to say if he had been a bit more willing to reconsider some of his views. Sadly, it's a rare scholar who is willing to admit that he's wrong.

[351]. "DN" stands for "double negation."

[352]. See Jerry Fodor's superb paper, "The present status of innateness controversy," in Fodor (1981b).

[353]. Kant (1787).

[354]. Quine (1970).

[355]. Frege (1956).

[356]. In this chapter, there are many places where, technically, I should have used quasi-quotation marks but where I instead used ordinary quotation marks or didn't use any quotation marks at all. I didn't want an excess of obscure symbolism to obscure matters of substance.

[357]. "LC" stands for "logical constant."

[358]. NOTE: The funny, square-shaped things flanking that expression are ***quasi-quotation marks.*** At the end of this chapter, it'll be explained what quasi-quotation marks are. For now, you may treat them as quotation marks—which, give or take some niceties, is what they are.

[359]. Wittgenstein (1922).

[360]. "AFT" is short for "argument for formal truth." AFT is a paraphrase, not a direct quotation.

[361]. Earlier we said that there are multiple answers to the question "what is ethics?" Our first answer was: "Ethics is the branch of philosophy that studies the categories peculiar to *normative* statements." We've just given a second answer to that question, namely: "it is the discipline that answers that attempts to determine which things are *intrinsically* good."

[362]. Aristotle held this view, or one like it. A version of it is defended in Chapter 22.

[363]. Examples are David Hume (1739, 1751), C.L. Stevenson (1937), R.M. Hare (1952), and Simon Blackburn (1984, 1998).

[364]. See Stevenson (1937).

[365]. A 19th century pontificator.

[366]. "HA" stands for "Hume's argument."

[367]. Why must he *know* himself to have such a desire? Why can't he just have it, without knowing that he has it? Because if he wanted to play soccer and, all of a sudden could do so, he obviously wouldn't be happy *unless he knew he could do so.* If he still thought his lungs would fill up with fluid when playing, he obviously couldn't be happy that he could play soccer without his lungs filling up with fluid. Since he'd believe, albeit wrongly, that his lungs *would* fill up, he'd be *unhappy* that (so he thought) he couldn't play soccer. So, I repeat, a *necessary* condition for Ben's *being happy that he can play soccer* is that he know himself have some desire that, in now being able to play soccer, he is able to gratify.

[368]. G.E. Moore made this insightful point in his (1903) book *Principia Ethica.*

[369]. Very technically speaking, it is a ***universal generalization,*** not a hypothetical. But it makes a statement *about* infinitely many hypotheticals, all having the form *if such and such is a Euclidean triangle, then its interior angles add up to 180°;* and although all those statements are true, not a single one of them has an antecedent that, the laws of physics being what they are, could even possibly be true.

[370]. In any case, it prohibits *given* the coordinative definitions that validate physics. See Frank (1949).

[371]. See Rachels (1986) for an excellent discussion of cultural relativism. Rachels makes some, not all, of the points just made.

[372]. See Chapter 24 for a discussion of utilitarianism. In Chapter 22, I defend a virtue-theoretic analysis of morality.

[373]. See Moore (1911).

[374]. The following argument is put forth by Philippa Foote in her paper "Morality as a system of hypothetical imperatives." I would like to point out that, before writing this chapter, I had a very illuminating discussion with my colleague Tony Ellis concerning this paper.

[375]. People want to be free and, therefore, unrestrained by anything, including morality. In many cases, the immoral acts that people perform for non-pragmatic reasons are acts of hostility towards others. It's clear that, in many such cases, what is motivating the agent is ill-will. But I suspect that, just as often, the primary motivation is the aforementioned desire not to be restrained and that this desire, though not itself an aggressive one, happened to find an outlet in an act of aggression.

[376]. Not necessarily a lot of self-consciousness. I believe that dogs and cats and many other mammals have more than enough and can, in fact, be happy and unhappy. Below I discuss why the oft-made claim that animals aren't self-conscious is a spurious one.

[377]. The theory advocated in this chapter relies heavily on Frankfurt's groundbreaking analyses of freedom and personhood. These are discussed in Chapter 15. It would be helpful to read that chapter before this one.

[378]. The fact that people (masochists) can derive pleasure from pain supports this analysis. Such people obviously value pain, showing that whether pain causes one unhappiness is, at least in part, a function of what one's values are.

[379]. The "corpus callosum" is the physical link between a person's two brain-halves. It was found that removal of the corpus callosum sometimes mitigated the symptoms of epilepsy. Roger Sperry (1913–1994) studied the psychological effects of this procedure and found that, quite literally, the right hand did not know what the left hand was doing. The clinical data strongly suggests that the post-operative body of a person whose corpus callosum had been removed is occupied by *two* people, albeit two people who are in very rapid, unspoken communication with each other. Because this isn't my area, I must refer the interested reader to Sperry's important work, which is easily found online.

[380]. The statement "mental event (or condition) M is irrational" can mean either (i) that M shows a lack of intelligence (or, in any case, a failure to deploy such intelligence as M has) *or* (ii) that M shows of a lack of sanity (or self-control).

Emotions *can* be irrational in sense (i). In other words, an emotion can embody a viewpoint that, given the data at one's disposal, one wouldn't have had *if* one had processed that data in an intelligent manner. But emotions aren't *categorically* irrational in sense (i). Many emotions embody no viewpoints that embody a failure to process the data in an effective manner.

Oftentimes, extreme emotions (operating in conjunction with other factors) cause psychological breakdowns; and often times they are indicative of psychological breakdown that have already occurred. So emotions *can* be irrational in sense (ii). But emotions aren't *always* irrational in that sense.

The notion that emotions are irrational in sense (ii) is very much at odds with the clinical data. A lack of emotion is an excellent indicator of an extreme disorganization of one's psyche.

Why is this, incidentally? Here is my hypothesis. The mind consists of a number of distinct modules. These modules must share information if the mind is to remain intact. Emotions facilitate the exchange of certain kinds of information between certain kinds of modules. (My reasons for holding this are outlined in Chapter 20.) When emotions are gone these intrapsychic communications are no longer possible. And the result is a break down in one's psychological integrity.

[381]. This isn't true of all animals, of course—only of those above a certain level. Where that cut-off is exactly, I do not know.

[382]. "MC" can be thought of as standing for "moral code."

[383]. This is assuming that there could even be language in a non-social world. If there couldn't, that would itself support my claim that the meaningfulness of the efforts of composers, physicists, authors, etc., are tethered to existence of cooperative social arrangements.

[384]. On the rare occasions when a psychopath seems to self-censure, he's using the tiny, poorly integrated fragment of conscience that he still has as a way of selling more lies to himself and others, so as to broaden the basis of his psychopathy. The psychopath isn't someone who has *no* conscience. He's someone who has one that functions only intermittently and that he doesn't identify with. And the little bit of conscience that the psychopath has becomes but another vehicle in his pursuit of amorality. "Nothing sells like sincerity." By having a bit of conscience, they amplify their powers of empathy, which is to be distinguished from sympathy, thereby perfecting their skills as operators.

(A colleague of mine, who wishes to remain anonymous, made me aware of the distinction between empathy and sympathy and of its relevance to the nature of psychopathy. This individual convincingly explained how psychopaths are, in fact, highly empathic—that, were they not hyper-empathic, they would be much less adept at manipulating people than they in fact are. Much of what I have to say on the topic of psychopathy I learned from conversations with this person, whose points have been borne out by my research into this topic and also by my, I am sorry to say, extensive personal experience with these fauxbrethren of ours.)

Assuming that you are not a psychopath (which you are probably not, since psychopaths don't value learning for its own sake), your conscience is exercising a constant pressure on you. That doesn't mean you can't perform bad acts. Non-psychopaths can be exquisitely evil, as we've already noted. It means that you are always under the watchful eye of a standard-bearer. Oftentimes people of conscience perform evil acts in order to rebel against their consciences. And sometimes, wanting to appease their own overly brutal consciences, non-psychopaths perform bad acts in order to provoke authority figures to punish them, as Freud pointed out in his paper "Criminals from a sense of guilt." But psychopaths don't labor under the watchful eye of a censorious superego.

Their condition may sound enviable. It is not. Because of their lack of conscience, they grossly under-perform. A psychopath who, given his intellectual level, could be a head of state or a great novelist, ends up being a two-bit con-man. Also, on the rare occasions when they're sincere about their condition, psychopaths describe their lives as being cold and empty. (Which, presumably, is why they are constantly in need of petty forms of stimulation, as Robert Hare points out in his (1999) book *Without Conscience*.) They describe themselves as having no selves at all, and wishing that they did have selves. Stories and movies about para-humans (e.g., robots, scarecrows) who wished they had souls are, I suspect, allegories about psychopathy.

In the movie, *Invasion of the Body Snatchers* soulless aliens take over people's bodies. After a given person's body had thus been "snatched," the resulting being was person-like but wasn't actually person. Such beings were superficially normal, were in some respects functional, and they had a calmness that mimicked a self-control that would ordinarily be indicative of depth of character. But their apparent lack of emotion was an expression, not of maturity and depth of character, but of immaturity and lack of character. Because they were so undeveloped, they couldn't integrate the emotions they had into the personae through which they represented themselves to the world. As a result, those personae were drained of any affect and thus bore a superficial resemblance to somebody who, unlike them, had confronted his emotions and mastered them. But unlike such a person, these pseudo-people were bubbling vats of emotion

that, when the slightest provocation caused the mask of sanity to fall off, burst out in fits of infantile hysteria. Because of the similarities between these pseudo-people, on the one hand, and psychopaths, on the other, I believe that *Invasion of the Body Snatchers* is a metaphorical representation of the process by which a person becomes psychopathized.

In that movie, on the rare occasions when the *actual* emotions of the pseudo-calm pseudo-people were revealed, they were base and infantile ones. (They really only had two emotional modalities: rage and satisfaction when the cause of their rage had been neutralized.) In fact, they were barely even emotions at all; they were too impersonal—too lacking roots in a coherent subject. I am thinking of the scene at the very end of the 1977 version of the movie, starring Donald Sutherland, where he points at his former comrade and inhumanly shrieks at her. Which is exactly how, when they were cornered, the psychopaths I've known behaved.

[385]. "MM" is short for "Mao morality."

[386]. See Hare (1999).

[387]. Kant's ethical views are put forth in his *Groundwork of the metaphysics of morals.*

[388]. "ACI" is short for "are there any categorical imperatives?"

[389]. "CI" stands for "categorical imperative."

[390]. See Chapter 19, Sections 6.0–6.3 for a discussion of the principle underlying the viewpoint now being put forth.

[391]. This argument is found in Kant's *Critique of Practical Reason;* it is hinted in his *Groundwork of the Metaphysics of Morals.*

[392]. "CI" stands for "categorical imperative."

[393]. The asterisk is meant to indicate that I'm not using the same wording as before.

[394]. Kant wrote a book titled *The Critique of Pure Reason* and one titled *The Critique of Practical Reason.* Few, if any, did as much to elucidate this distinction as Kant. Aristotle was also aware of the distinction and did much to elucidate it. In my judgment, Kant and Aristotle are, ultimately, the best ethicists of all time. (I doubt I'm alone in this judgment.) It's no accident that the ethical systems of these two ethicists, unlike those of most (if not all) others, were efforts to validate the assumption that rationality and morality are, in some way or other, coincident. Harry Frankfurt is another such ethicist (and, in my judgment, his work is, partly for that reason, uniquely valuable).

[395]. "NSG" is short of "no self-government."

[396]. This section presupposes that dogs aren't rational. If they are, then pretend that I'm talking about some other, less gifted sort of animal.

[397]. It seems wrong to describe animals as "irrational." The way in which a chipmunk's mind falls short of rationality isn't comparable to the way done so by the mind of an otherwise rational human being in the grips of a psychotic break. The latter's basic condition is one of rationality but his current condition is not one of rationality. But a chipmunk's failure to be rational *is* its basic condition, and isn't one of illness. The otherwise rational, but currently psychotic, person is said to be "irrational." We'd make that word ambiguous were we to use it to describe the chipmunk. So I'll describe chipmunks, and other such things, as "sub-rational."

[398] Psychoanalysts say that one often engages in "intellectualization" so as to escape from what is really bothering one and so as to facilitate one's ability to continue to deal with it in a neurotic, self-defeating way.

[399]. Rachels (1986) makes many, but by no means all, of the points made in this chapter.

[400]. Also, if psychological hedonism is correct, a person isn't even capable of telling people that ethical hedonism is correct unless he believes that by doing so he is maximizing the amount of pleasure he will

experience. Given how unlikely it is that a given person has this belief, the words of a psychological hedonist who advocates ethical hedonism deserve no credence.

[401]. This is known as "the doctrine of double effect." St. Thomas Aquinas (1225–1274) put forth this doctrine in his *Summa Theologica.*

[402]. Many would say that it is with the intention of experiencing pleasure that people have sexual intercourse and perform other such carnal acts. Though this is certainly a large part of the truth, sexual activity isn't in the same category as narcotics use. People's sexual practices are replete with beliefs as to what is of value. Even sex acts that we see as particularly anti-social and perverse are, I suspect, expressions of values of a kind. It may even be that what offends people about certain sexual practices that are customarily described as "perverse" or "sick" is precisely that they embody values of a kind that people find inimical to their interests. (A lot of supposed sexual pseudo-immorality is actually hyper-morality or anti-morality—anti-morality is a form of hyper-morality. The only people who become anti-moral, as opposed to simply being amoral, are those who are overburdened by conscience.) I also suspect that what draws people to such acts is precisely that they see them as expressive of certain values. So while it's an obvious that much of what motivates people to engage in such activity is the desire for pleasure, that's not all that such activity is about. And because sexual activity, even (especially) in its supposedly less laudable forms, is psychologically so multi-dimensional, it shouldn't be co-categorized, at least not in this context, with narcotics use.

[403]. Except insofar as having a good time *qua* having a good time has value. It would be unhealthy for a person to permit himself no pleasures other than ones that he acquired in the course of activities that he valued and that he didn't perform with the intention of feeling good.

[404]. Notice that the words we use to refer to positive psychological conditions incorporate rather subtle judgments as to the moral characteristics of those conditions. For example, what Smith feels in doing cocaine is "pleasure." It isn't "joy," let alone "happiness." Joy is

what Beethoven felt. If the reason that one feels good in that what one is doing has value, what one experiences is "joy." If one feels good, but not because what one is doing has value, one is experiencing "pleasure."

Joy and pleasure are not mutually exclusive. There's no denying that pleasure is an important part of joy. Valuable activities are often pleasurable ones. (And it is usually *in virtue of its being valuable* that one's work brings one pleasure. It isn't usually because, for some random reason, what one is doing happens to set off some random causal structure that eventuated in their experiencing pleasure.)

But there can be pleasure without joy. There can, in other words, be *mere* pleasure. And there are very few people, if any, who don't at some level regard a life dedicated to *mere* pleasure as a good one. There are very few true hedonists.

I suspect that self-described hedonists are really people who sense that there's something wrong with conventional values—something in them that is inimical to human welfare—but who don't realize that their discomfort with conventional values is rooted, not in their having *no* values or only trivial values (e.g., "have a good time"), but in their having very robust *contrary* values. Not knowing how to describe their views, they say that they're "hedonists" (or "nihilists" or some such), when, in fact, they're nothing of the sort. To the extent that one is *actually* a hedonist, one doesn't feel the need to describe oneself as a hedonist. People don't label their practices unless they're trying to justify them, and a true hedonist ipso facto doesn't feel the need to justify anything.

[405]. See John Stuart Mill's book *Utilitarianism.*

[406]. Why the "so it could be argued"? Because I don't think that the world would plunge into a permanent war of all nations against all nations if there were no international laws (or law-like protocols) prohibiting nations from intervening in the affairs of other nations. Nations are allowed to do whatever they want. The laws that govern nations, supposing that there even are such laws, aren't nearly as robust as the laws that govern individuals within nations. What prevents nations from invading other nations is that, for reasons having nothing to do with the

existence of international law, it would not be in their strategic selfinterest to do so.

[407]. See the previous chapter.

[408]. See David Ross's classic defense of deontology *The Right and the Good.*

[409]. My understanding of the topics discussed in this section benefited in a very genuine way from conversations with Anthony Ellis.

[410]. In his (1991) article "Consequentialism," Philip Pettit makes a strong case that deontology is an incoherent doctrine.

Nonetheless, I am open on this issue. But, speaking emotionally as opposed to discursively, my sympathies are with the consequentialist. In my personal experience, which is obviously of limited probative value, many a coward justifies his failure to do the right thing by saying that he doesn't want to "play God" or by saying that, even though the circumstances clearly demand it, it's not "his place" to question accepted institutional protocols.

I've also found that where there is abject cowardice, there is also a good deal of ill will. (I suspect that cowards are ashamed of their condition and projectively lash out when given the chance.)

Finally, I have found there to be a direct correlation between, on the one hand, how often the word "fair" (as in "that wouldn't be fair, would it?") occurs in a person's discourse and, on the other hand, how ill-willed that person is. I am proud to say that, in my 350 page dissertation on the ethical basis of morality, the word "fair" doesn't occur once, and the word "justice" occurs *only* once—and it occurs in the context of a quotation.

[411]. For the record, I don't think this is true. I think that, at the very least, it's an exaggeration. I think that genuinely well-intentioned people wouldn't have created the sorts of political structures that made Stalin, et al. possible.

[412]. "OG" is short for "omnipotent God."

[413]. "AG" is short for "Aquinas on God."

[414]. This argument and the next are found in Aquinas' *Summa Theologica.*

[415]. So is its Latin equivalent. (Aquinas wrote in Latin.)

[416]. There are other problems with AG_1. For example: It starts off by saying that every event has a predecessor, and ends by saying that there is a first event (i.e., that *not* every event has a predecessor).

[417]. What does it mean to say that it can be parsed in more than one way? Sentences aren't strings of words. They're strings of phrases. A phrase is the largest grammatically cohesive unit that is smaller than a sentence. To say that a sentence can be parsed in different ways is to say that there are different ways of grouping the words composing it into phrases.

[418]. "Credo quia absurdam"—I believe because it's absurd. The medievals said this. They were right—after a fashion. I learned of this expression from reading Freud's (1933) book *Civilization and its discontents.*

[419]. St. Thomas Aquinas (1225–1274) accepted the argument about to be presented. This would obviously seem to be inconsistent with my statement that religious people reject it. But—and I say this with all due to respect to devout Catholics—my suspicion is that, ultimately, he was a philosopher who, living when he did, had to express his philosophical views in religion-friendly terms. True religion is fundamentalism, and fundamentalists don't accept the Euthyphro argument. (Given their views, they are right not to accept that argument.)

[420]. "AF" stands for "advocate of the Free-Will defense."

[421]. This point was made by Bertrand Russell (1920).

[422]. "AG" is short for "argument for God's existence."

[423]. "MCS" stands for "meaning of CS."

[424]. Nathan Salmon (2005: Chapters 1–3) argues that it can meaningfully and *correctly* be said of nonexistent entities that they exist—and also don't.

[425]. Russell rejects the ontological argument for much the reasons just given. The argument just given is a refinement of one that he put forth. Kant too rejected the ontological argument. The problem with it, he said,

is that it falsely assumes that "existence is a predicate." It isn't clear exactly what he meant by this, but what he had in mind was obviously similar to what we, following Russell, have been saying.

[426]. Dworkin (1963).

[427]. In what remains of this chapter, I am inconsistent and sloppy in my use of quotation marks and quasi-quotation marks. Technically-minded readers are urged not to put too much stock in the way in which, in this chapter, I use such devices. In his compact and beautiful (1941) book *Mathematical Logic,* Quine is exceptionally mindful of the sometimes subtle distinction between quotation and quasi-quotation. (Quine invented quasi-quotation marks, and it was in that very work of Quine's that he introduced them to the world.)

[428]. In medieval logic, it did have a meaning. But medieval logic wasn't very good, and the distinction that it made between negations and opposites—or "contraries," as they were sometimes called—isn't very useful. So we'll ignore it.

[429]. Given that "because" operates on ordered pairs of sentence, it is *de rigueur,* I believe, to hold that "or" and "and" do so as well. It may be that "grass is green and snow is white" has the same literal meaning as "snow is white and grass is green." But if, on that account, we say that "and" operates on unordered pairs, we are saying by implication that a knowledge of logic—of the logical properties of conjunction—is embedded in our syntactic conventions. And that seems quite false. A mastery of logic isn't a prerequisite for a mastery of English syntax. Since it would be unreasonable to assume that a knowledge of the fact that conjunction, unlike causal dependence, is a symmetrical relation, we must assume that, so far as one's knowledge of English syntax and semantics is concerned, "and" and "because" are comparable. It's clear that anyone who speaks English knows that "because" is an order-sensitive operator. We must therefore assume, I believe, that *qua* speaker of English, one rightly believes "and" to be an order-sensitive operator—that this belief, though obviously not in all cases discursively held, is embedded in one's linguistic competence.

[430]. See Strawson (1950), Barwise (1983), and Barwise and Perry (1999).

[431]. "It is necessarily the case that" is thus the dual of "it is possible that." Given any operator O, any operator O* is the dual of O if, given an arbitrary sentence S, O*S iff ~O~S. Thus, "for all x" is the dual of "there exists some x," since "for all x, x is tall" is equivalent with "it is not the case that, for some x, it is not the case that x is tall."

Exercise: prove that duality is a symmetrical relationship. In other words, prove that, if O is the dual of O*, O* is the dual of O.

Second exercise: Identify two operators O and O* such that each is the dual of the other. If you can't find two such operators, invent them.

[432]. Technically, this section should be called "some rudiments of symbolic logic"—the word "formal" isn't quite appropriate. The reason is that Modus Ponens, Modus Tollens, and other mainstays of "formal logic" are not themselves formally true. First of all, "((P and (P→Q))→Q" isn't itself a statement at all. It's a statement form. The corresponding statement is: "Given any propositions P and Q, (P and (P→Q))→Q" , which isn't formally true, as it has the same form as "Given no propositions P and Q, (P and (P→Q))→Q."

[433]. Peter Strawson makes this point in his landmark (1950) paper "On referring."

[434]. The scope of the box-operator only includes the P. So 12 says that, if P is necessary, it follows that Q entails P. 12 does *not* say that *it's necessary that P entails that Q entails that P.* The italicized principles is false, being equivalent to the obvious falsehood that every statement is entailed by every statement.

[435]. In "The logic of what might have been," in Salmon (2005), Nathan Salmon argues that S5 is false.

[436]. "MT" is short for "model-theoretic."

[437]. Actually, P and Q can't really be sentences if this is to be true. Nor can they be sentence-tokens. It must be supposed that they express *propositions.* But since formal entailment is a relationship between expressions, not propositions, it isn't an option for formal logic to construe modus ponens (if P and (P→Q), then Q) as being to the effect that: for any *propositions* P and Q, (if P and (P→Q), then Q). The formal

logician must find expressions that, if the substituents of P, Q, etc. validate that statement. This is very hard to do, which exposes an incoherence that lies at the center of formal logic.

[438]. Actually, P and Q can't really be sentences if this is to be true. Nor can they be sentence-tokens. It must be supposed that they express *propositions.* But since formal entailment is a relationship between expressions, not propositions, it isn't an option for formal logic to construe modus ponens (if P and (P→Q), then Q) as being to the effect that: for any *propositions* P and Q, (if P and (P→Q), then Q). The formal logician must find expressions that, if the substituents of P, Q, etc., validate that statement. This is very hard to do, which exposes an incoherence that lies at the center of formal logic.

[439]. See Quine (1970).

[440]. Hempel (1945a, 1945b) holds this.

[441]. In his (1976) paper "Can there be vague objects?" Gareth Evans says that, according to a viewpoint that he rejects, "rather than being a deficiency in our mode of describing the world, [vagueness] would be a necessary feature of any true description of it." For reasons of professional ethics, I wish to bring the reader's attention to the (unintentional, at least at the level of consciousness) similarity between my wording and Evans'.

[442]. Contrary to what Quine alleged, Pierre Duhem (1861–1916) did not put forth the so-called "Duhem-Quine" thesis.

[443]. Though I believe there to be a propositional analogue of it. See Chapter 3.

[444]. There isn't really such a thing as *the* logical form of a sentence. What it is that counts as a clarification of a sentence's meaning depends in part on what one's starting-assumptions are. For example, if one takes the concepts *point, plane,* and so on, as primitive, the sentence "x is the class of points equidistant from a given point on a plane" might give the logical form of "x is a circle." If one doesn't take those concepts as primitive, it might abjectly fail to do so.

Also, given some language L that is expressively more powerful than English, the L-sentence that would give the logical form of "snow is

white" would not be a direct translation of the English sentence that would do so. So it could be sentence that the logical form of "snow is white" *relative to L* isn't identical with the logical form of "snow is white" relative to English. (That said, I'm not entirely convinced that logical form is relative in this way.)

[445]. For example, Peter Strawson in his otherwise brilliant (1950) paper "On referring."

[446]. The sentence "this is radically absurd" occurs in Russell's (1903) book *Principles of Mathematics* (PM). (It occurs in the chapter on Newton's laws of motion.) I rather fancy this sentence, and thus use it from time to time. Ironically, the statement that Russell describes as "radically absurd" in PM is, as I recall, quite a defensible one.

[447]. "SS" is short for "Slingshot."

[448]. Advocates of this argument say that the true sentences refer to *The True*, and that false sentences refer to *The False.* They don't say what these terms mean. In an effort to free their argument of needless obscurities and indefensible assumptions, I've taken the liberty of assuming that they refer to the properties of being true and false, respectively.

[449]. Give or take some irrelevant details, this argument was put forth by Frege in his landmark paper *On sense and reference.*

[450]. See his paper "True to the facts," in the (1984) collection of his papers *Essays on Truth and Interpretation.*

[451]. Oxford University Press, 2005.

[452]. In what follows, I use actual quotation marks where I should, stricty speaking, use quasi-quoatatoin marks.

[453]. To my knowledge, this term was coined by John Searle. It is the title of his excellent (1969) book *Speech acts.*

[454]. Examples of universal quantifiers are "for all x," "given any x," "any object," "all objects," and so on. (In general, a universal quantifier is an expression E such that, in virtue of having the form ‹...E...›, a sentence expresses a proposition to the effect that all members of some one class are members of some other class. Thus, "for all x, if x is a dog, then x is a mammal" is to the effect that every member of the class of dogs is a

member of the class of mammals.